International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 25, No. 8 (2011) 1173–1195
© World Scientific Publishing Company
DOI: 10.1142/S0218001411009093

MEASURING THE PERFORMANCE OF ORDINAL CLASSIFICATION

JAIME S. CARDOSO* and RICARDO SOUSA†

INESC Porto, Faculdade de Engenharia
Universidade do Porto, Campus da FEUP
Rua Dr. Roberto Frias, n 378
4200-465 Porto, Portugal
**jaime.cardoso@inescporto.pt*
†rsousa@inescporto.pt

Ordinal classification is a form of multiclass classification for which there is an inherent order between the classes, but not a meaningful numeric difference between them. The performance of such classifiers is usually assessed by measures appropriate for nominal classes or for regression. Unfortunately, these do not account for the true dimension of the error.

The goal of this work is to show that existing measures for evaluating ordinal classification models suffer from a number of important shortcomings. For this reason, we propose an alternative measure defined directly in the confusion matrix. An error coefficient appropriate for ordinal data should capture how much the result diverges from the ideal prediction and how *"inconsistent"* the classifier is in regard to the relative order of the classes. The proposed coefficient results from the observation that the performance yielded by the Misclassification Error Rate coefficient is the benefit of the path along the diagonal of the confusion matrix. We carry out an experimental study which confirms the usefulness of the novel metric.

Keywords: Classification; ordinal data; evaluation measures; performance; classification accuracy.

1. Introduction

In many real life problems humans are called to compare or rank items or objects in order to make the most appropriate choice for a specific goal. Think for example of choosing a song to listen, buying clothes, ordering a dish in a restaurant, etc. Other applications include stock trading support systems, where one wants to predict, for instance, whether to buy, keep or sell a stock, and biomedical classification problems, where frequently the classes are ordered. As a consequence, the demand for intelligent systems capable of representing and processing this information also increases. In research areas like decision making, preference modeling, fuzzy modeling, statistics and machine learning, scientists have proposed various ways to characterize this human behavior with mathematical models.

Mainly two learning settings can be distinguished for modeling preference information: ordinal classification models and pairwise preference models, both dependent on the concept of an underlying ranking function.[10,23] We concentrate only on the ordinal classification setting, where categories typically correspond to quotations or linguist terms — varying from "very bad" to "excellent," for example — that express a difference in correctness, quality, beauty or any other characteristic of the analyzed objects.[23] A concrete example would be the grading of a customer credit profile in the scale *Excellent* $\succ$ *Good* $\succ$ *Fair* $\succ$ *Poor* or grading a student in a similar scale, where $\succ$ is the order relation.

One of the first works on classification methods for ordinal data dates from McCullagh[14] where a regression model was developed incorporating ordinal information on the data. An extension of this work is presented in Ref. 21 through the generalization of the additive model[8] by incorporating nonparametric terms. Frank and Hall[6] introduced a simple process to explore the ordinal class information by using conventional binary classifiers. In Ref. 19 a generalised formulation for the SVM was introduced for ordinal data. More recently, Ref. 12 proposed a cascade classification technique encompassing a decision tree classifier and a model tree algorithm. In Refs. 4 and 15 two new methods were present towards ordinal classification. In Ref. 4 a new reduction technique is used allowing to solve the problem of ordinal classification using a single binary classifier. In Ref. 15 the class order relation is taken into account by imposing an unimodal distribution to the class *a posteriori* probabilities. An extension of this technique on All-at-Once SVM is performed in Ref. 16.

In supervised classification problems with ordered classes, it is common to assess the performance of the classifier using measures more appropriate for nominal classes, regression problems or preference learning.[1,7] Baccianella[1] addresses the adaptation of existing measures (Mean Absolute Error) to unbalanced data, while Gaudette[7] compares existing measures concluding that Mean Absolute Error and Mean Square Error are the best performance metrics. Other strategies encompass the use of rank order measures[13,22] or the adaptation of the ROC curve.[24] However, the application of these measures faces difficulties in the context of ordinal classification, as we will show next.

In this manuscript, our main goal is to propose a new metric specifically adapted to ordinal data classification problems, problems endowed with a natural order among classes. We argue that standard metrics do not adequately take into account all the information in the assessment process. We also claim that an error coefficient appropriate for ordinal data should capture how much the result diverges from the ideal prediction and how "inconsistent" the classifier is in regard to the relative order of the classes. This "inconsistency" results from discordant results in the relative order given by the classifier and the true relative class order. For this reason, in this work, we propose an alternative measure defined directly in the confusion matrix.

1.1. *Common evaluation measures for ordinal classification*

Very often, every misclassification is considered equally costly and the Misclassification Error Rate (MER) is used. Two other measures also usually applied are the Mean Absolute Error (MAE) and the Mean Square Error (MSE). Both MAE and MSE address the problem as a regression task, i.e. the performance of a classifier is assessed in a dataset $\mathcal{O}$ through

$$\frac{1}{N}\sum_{\mathbf{x}\in\mathcal{O}}|g(\mathcal{C}_{\mathbf{x}})-g(\hat{\mathcal{C}}_{\mathbf{x}})|$$

and

$$\frac{1}{N}\sum_{\mathbf{x}\in\mathcal{O}}(g(\mathcal{C}_{\mathbf{x}})-g(\hat{\mathcal{C}}_{\mathbf{x}}))^2$$

respectively, where $g(.)$ corresponds to the number assigned to a class, $N=\text{card}(\mathcal{O})$, and $\mathcal{C}_{\mathbf{x}}$ and $\hat{\mathcal{C}}_{\mathbf{x}}$ are the true and estimated classes. However, this assignment is arbitrary and the numbers chosen to represent the existing classes will evidently influence the performance measurement given by MAE or MSE. A clear improvement on these measures would be to define them directly from the confusion matrix CM (a table with the true class in rows and the predicted class in columns, with each entry $n_{r,c}$ representing the number of points from the rth class predicted as being from cth class):

$$\text{MAE}=\frac{1}{N}\sum_{r=1}^{K}\sum_{c=1}^{K}n_{r,c}|r-c|$$

$$\text{MSE}=\frac{1}{N}\sum_{r=1}^{K}\sum_{c=1}^{K}n_{r,c}(r-c)^2$$

where K is the number of classes. We will always assume that the ordering of the columns and rows of the CM is the same as the ordering of the classes. This procedure makes MAE and MSE independent of the numbers or labels chosen to represent the classes. To a certain degree, these two measures are better than MER because they take values which increase with the absolute differences between "true" and "predicted" class numbers and so the misclassifications are not taken as equally costly. Still, these measures do present undesired behavior, as we will show later.

In order to avoid the influence of the numbers chosen to represent the classes on the performance assessment, it has been argued that one should only look at the order relation between "true" and "predicted" class numbers. The use of Spearman's rank correlation coefficient, R_s, and specially Kendall's tau-b, τ_b, is a step in that direction.[20,11] For instance, in order to compute R_s, we start by defining two rank vectors of length N which are associated with the variables $g(\mathcal{C})$ and $g(\hat{\mathcal{C}})$. There will be many examples in the dataset with common values for those variables; for these cases average ranks are used. If $\boldsymbol{p}$ and $\boldsymbol{q}$ represent the two rank vectors, then

$R_s = \frac{\sum(p_i-\bar{p})(q_i-\bar{q})}{\sqrt{\sum(p_i-\bar{p})^2\sum(q_i-\bar{q})^2}}$. As we can see, Spearman's coefficient is still dependent on the values chosen for the ranks representing the classes and so it is not completely appropriate to measure the performance of ordinal data classifiers. More importantly, R_s loses information about the absolute value of the classes.

Kendall's coefficient τ_b has been advocated as a better measure for ordinal variables because it is independent of the values used to represent classes.[11] Its robustness is achieved by working directly on the set of pairs corresponding to different observations. To define τ_b, start with the two N-point vectors, associated with the true and predicted classes, $\mathcal{C}_x$ and $\hat{\mathcal{C}}_x$, and consider all $\frac{1}{2}N(N-1)$ pairs of data points. Before proceeding, some definitions are required.[17]

Definition 1 (Concordant Pair). We call a pair (i, j) *concordant*, c, if the relative ordering of the true classes $\mathcal{C}_{x_i}$ and $\mathcal{C}_{x_j}$ is the same as the relative ordering of the predicted classes $\hat{\mathcal{C}}_{x_i}$ and $\hat{\mathcal{C}}_{x_j}$.

Definition 2 (Discordant Pair). We call a pair *discordant*, d, if the relative ordering of the true classes is opposite from the relative ordering of the predicted classes.

Definition 3 (Pair Ties). If there is a tie in either the true or predicted classes, then we do not call the pair either concordant or discordant. However, different concepts applies to different types of ties.

extra true pair: If the tie is in the true classes, we will call the pair an *extra true pair*, e_t.

extra predicted pair: If the tie is in the predicted class, we will call the pair an *extra predicted pair*, e_p.

ignore pair: If the tie is both on the true and the predicted classes, we ignore the pair.

The τ_b coefficient can be computed as

$$\tau_b = \frac{c-d}{\sqrt{c+d+e_t}\sqrt{c+d+e_p}}$$

where c refers to concordant pairs and d for discordant pairs. The τ_b coefficient attains its highest value, 1, when both sequences agree completely, and -1 when the two sequences totally disagree. However, the source of robustness is probably the source of its main limitation: by working only with the relative order of elements, it loses information about the absolute prediction for a given observation, making the coefficient more suitable for assessing preference learning[18] rather than ordinal data classification.

In the same line, the coefficient r_{int} was recently introduced, taking into account the expected high number of ties in the values to be compared.[15] In fact, the variables $\mathcal{C}$ and $\hat{\mathcal{C}}$ are two special ordinal variables because, as there are usually very few classes compared to the number of observations, these variables will take many tied values (most of them, in fact). Nevertheless, r_{int} is sufficiently general and, if there are no tied values, it can still be applied as it is. Like τ_b, r_{int} assumes that the only thing that matters is the order relation between such values, which is the same as the order relation between the classes. This coefficient takes values in $[-1, 1]$.

Note that MER and MAE are indices of dissimilarity while R_s, τ_b and r_{int} are indices of similarity. It is important to remark right now a limitation of MAE (and MSE). Start by noticing that the range of possible values for MAE is an upper-unbounded interval. Nevertheless, it is fair to compare MAE results in two different applications with a different number of observations, N, since MAE is properly normalized by N. However, if the applications involve a different number of classes, K, it is not clear how to compare the performance obtained in the two settings.

2. A Preliminary Comparison of the Merits of Existing Metrics

A major difficulty in the design of a new classification performance coefficient lies in the difficulty in demonstrating that the coefficient captures adequately the performance of the classification algorithms. In a first test to check the adequacy of the coefficients discussed in the previous section, we created synthetic classification results and compared the values given by the coefficients with the expected measured performance. The performance of any classification algorithm is conveniently summarized in the CM and any of the coefficients presented in the previous section can be computed directly from it. Suppose that four classifiers A, B, C and D produce the following the CMs ($K = 4$, $N = 13$) in a certain task:

$$\text{CM}(A) = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad \text{CM}(B) = \begin{bmatrix} 0 & 4 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

$$\text{CM}(C) = \begin{bmatrix} 0 & 0 & 4 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad \text{CM}(D) = \begin{bmatrix} 0 & 4 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

One would expect that a valid measure of performance would output for classifier A a perfect performance, for classifier B an inferior performance and for classifier C a performance below B's performance. Table 1 presents the results for the different coefficients.

Note that R_s, τ_b and r_{int} were unable to detect any performance difference between classifiers A and B; that results from the fact that they only measure relative values. We can also conclude that, in this context, $1 - R_s$, $1 - \tau_b$ and $1 - r_{\text{int}}$ do not constitute metrics since they do not satisfy the identity of indiscernible property ($d(x, y) = 0$ if and only if $x = y$). The MER coefficient was unable to differentiate classifiers B and C; note that, since classes are ordered, it is worse to predict points

Table 1. Results for the preliminary comparison, with $\beta_1 = \frac{0.25}{N(K-1)}$ and $\beta_2 = \frac{0.75}{N(K-1)}$. Coefficients $\mathrm{OC}^1_{\beta_1}$ and $\mathrm{OC}^1_{\beta_2}$ will be introduced later in the text.

Classifier	MER	MAE	R_s	τ_b	r_{int}	$\mathrm{OC}^1_{\beta_1}$	$\mathrm{OC}^1_{\beta_2}$
A	0.0	0.0	1.0	1.0	1.0	0.0	0.0
B	0.77	0.77	1.0	1.0	1.0	0.50	0.63
C	0.77	1.08	0.79	0.75	0.80	0.61	0.78
D	0.77	0.77	0.24	0.11	0.53	0.65	0.72

from class $\mathcal{C}_1$ to belong to class $\mathcal{C}_3$ rather than to predict them to be from class $\mathcal{C}_2$. The MAE coefficient (MSE would present the same behavior) was unable to differentiate classifiers B and D; note that classifier B was more consistent than classifier D in the sense that the relative order of the predicted classes coincides with the true order of the classes.

Finally, one can discuss the relative merit of C and D classifiers. If the ranking-based error is more relevant than the instance-based error then C should be preferred over D since the relative evaluation of C is consistent with the correct classification. When the instance-based error is prominent over the ranking error then one should prefer classifier D. We will return to this point later.

3. The Ordinal Classification Index

Nominal data classification analyzes each item in isolation and the closeness of the predicted assignment with respect to the exact one is the most relevant criterion. Ranking, which is an aggregate evaluation task, is instead totally focused on respecting the ordering of items, not considering the actual values assigned to them. When applied to ordinal classification, a drawback of any pairwise criteria, such as Kendall's coefficient, is that it does not allow example dependent evaluation.

At the heart of the proposed measure is the incorporation of a ranking-based component to an instance-based evaluation of ordinal classification. Nevertheless, the new metric is still applicable to the evaluation of single points.

An appropriate error coefficient for ordinal data should capture how much the result diverges from the ideal prediction and how much "inconsistent" the classifier is in regard to the relative order of the instances. We propose to define a metric directly in the CM, capturing these two sources of errors.

For this we adopt the following definition of *nondiscordant pair of points*:

Definition 4 (Non-Discordant Pairs). A pair of points x_i and x_j is called *nondiscordant* if the relative order of the true classes $\mathcal{C}_{x_i}$ and $\mathcal{C}_{x_j}$ is not opposite to the relative order of the predicted classes $\hat{\mathcal{C}}_{x_i}$ and $\hat{\mathcal{C}}_{x_j}$ (if there is a tie in either the true or predicted classes, or both, the pair is still *nondiscordant*).

In the CM, Definition 4 is translated into

$$\operatorname{sign}((r_{x_i} - r_{x_j}) \times (c_{x_i} - c_{x_j})) \geq 0 \tag{5}$$

where r_{x_i} and c_{x_i} are the row and column in the CM corresponding to example x_i, respectively. Finally, define a path in the CM as a sequence of entries where two consecutive entries in the path are 8-adjacent neighbors. The benefit corresponding to a path is the sum of the values of the entries in the path. In fact, it is useful to consider a graph associated with the CM, where each entry of the matrix corresponds to a vertex and there is an edge connecting vertices corresponding to adjacent entries.

The coefficient to be proposed results from the observation that the performance yielded by the MER coefficient is the benefit of the path along the diagonal of the CM. The MER coefficient only counts the pairs in the main diagonal of the CM to measure the performance; any deviation from the main diagonal is strictly forbidden — see Fig. 1(a).

A more relaxed coefficient can be defined by allowing the pairs to deviate from the diagonal, while staying *nondiscordant*. Therefore, we allow all pairs forming a consistent path from (1,1) to (K, K) — see Fig. 1(b). A path is said to be consistent if every pair of nodes in the path is nondiscordant. It is trivial to verify that any monotonous path (a path where the row and column indices do not decrease when walking from (1,1) to (K, K)) is consistent. The consistency of the classifier is therefore taken into account by valuing only the *nondiscordant* subsets of entries. Still, it is not enough to select the consistent path with the maximum benefit.

One should also penalize the deviation of the path from the main diagonal. We propose then to find the consistent path from (1,1) to (K, K) that maximizes the sum of the entries in the path and minimizes a measure of the deviation from the main diagonal. We propose the ordinal classification index OC_β to take the shape

$$\mathrm{OC}_\beta = \min\left\{\left(1 - \frac{1}{N}\mathrm{benefit}(\mathrm{path})\right) + \beta(\mathrm{penalty}(\mathrm{path})\right)\right\}$$

where the minimization is performed over the set of all consistent paths from (1,1) to (K, K) and $\beta \geq 0$. Tentative solutions for the penalty of the path include the excess

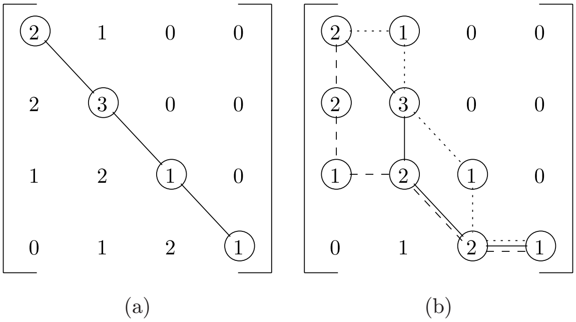
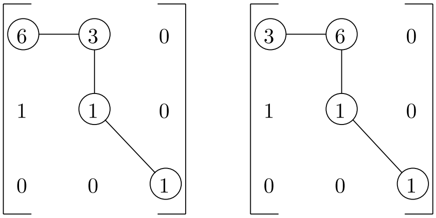


Fig. 1. Consistent paths over the CM. (a) An illustration of the benefit of the MER coefficient as the sum of the entries in the main diagonal of the CM. The MER coefficient results as $\frac{N-\mathrm{benefit}}{N}$. (b) Some examples of consistent paths; any pairs of observation contributing to the entries in a consistent path are *nondiscordant*. The benefit of a path is the sum of the entries in the path.



(a) (b)

Fig. 2. The two paths (a) and (b) would have the same penalization using the length, the maximum distance to the main diagonal or the area to select the cost; however, path (a) should be preferred over path (b).

on the length of the path over the minimum possible length (penalty(path) = length(path) − K), the maximum distance of the path to the main diagonal or the area between the path and the main diagonal. However, it is intuitive that these terms do not meet the required properties. In Figs. 2(a) and 2(b) we present two paths that would experience the same penalization under a measure based on the length of the path, the maximum distance to the main diagonal or the area of the path; however, it should be consensual that the CM in Fig. 2(a) represents a better performance than the CM in Fig. 2(b).

A penalization term suggested by the expressions of MAE and MSE is based on penalizing each vertex of the path by its "distance" to the main diagonal, obtaining

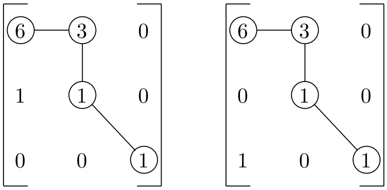
$$\mathrm{OC}'^{\gamma}_{\beta} = \min\left\{\left(1 - \frac{1}{N}\sum_{(r,c)\in\text{path}} n_{r,c}\right) + \beta\sum_{(r,c)\in\text{path}} n_{r,c}|r-c|^{\gamma}\right\} \tag{6}$$

where $\gamma > 1$. It is clear that $\mathrm{OC}'^{\gamma}_{\beta}$ is always non-negative, as the two terms in Eq. (6) are both non-negative; $\mathrm{OC}'^{\gamma}_{\beta}$ is also not superior to 1 as $\mathrm{OC}'^{\gamma}_{\beta}$ is always not superior to the cost over the main diagonal, where the path penalty is zero. It is also easy to conclude that if $\beta \geq 1$ then $\mathrm{OC}'^{\gamma}_{\beta}$ will equal the misclassification error (MER): since any deviation from the main diagonal will incur in a cost not inferior to 1, the optimal path is always over the main diagonal.

Nevertheless, this setting is still unsatisfactory; incorporating in the objective function only terms measuring the quality of the path does not capture differences in performance due to the leftover entries — see Figs. 3(a) and 3(b). One needs to also penalize the "dispersion" of the values from the main diagonal.

A first tentative solution is to add an additional term $\beta_2(\sum_{\forall(r,c)} n_{r,c}|r-c|^{\gamma})^{1/\gamma}$ to the objective function penalizing such dispersion of the data. This approach suffers from the disadvantages of adding a further parameter whose value needs to be selected and of changing the range of possible values for $\mathrm{OC}^{\gamma}_{\beta}$ from $[0,1]$ to an upper-unbounded interval.

Therefore, we propose to change the definition (6) by normalizing the benefit of the path not by N but by $N + M$, where $M = (\sum_{\forall(r,c)} n_{r,c}|r-c|^{\gamma})^{1/\gamma}$ is a measure of



(a) (b)

Fig. 3. The performance represented by CM in (a) should be better than the performance represented by CM in (b).

the dispersion of the data in the CM:

$$\mathrm{OC}_{\beta}^{\gamma} = \min\left\{1 - \frac{\sum_{(r,c)\in\text{path}} n_{r,c}}{N + \left(\sum_{\forall(r,c)} n_{r,c}|r-c|^{\gamma}\right)^{1/\gamma}} + \beta \sum_{(r,c)\in\text{path}} n_{r,c}|r-c|^{\gamma}\right\} \tag{7}$$

Note that M can be interpreted as the Minkowski distance between the two vectors used to build the CM. The parameter β controls the tradeoff between the relevance of the ranking-based component and the instance based evaluation. Small values for β will favor ranking over "absolute" classification; high values for β will do the opposite. In Table 1 we present the results for two different values of β. The only difference is the relative merit of classifiers C and D, in accordance with the preceding discussion.

3.1. *The ordinal classification index — general formulation*

Thus far, the consistency was valued by working only with nondiscordant pairs of points. The feasible paths were constrained under the set of consistent paths. A standard procedure in optimization is to replace a constraint by a penalty term in the goal function. Assume now we extend the set of feasible paths to the set of paths starting in (1,1) and ending in (K, K). Note also that there is always one of such paths going through all the entries in the CM. One can generalize the framework over this set of paths, penalizing now not only the deviation of the path from the main diagonal, but also the inconsistency of the path. One can therefore add an additional penalizing term to the definition of the index, capturing this undesirable attribute. An intuitive penalization term is the number of discordant pairs of vertices in the path, $N_{\text{disc\_pos}}$ (see (5)):

$$\begin{aligned}\mathrm{OC}_{\beta_1;\beta_2}^{\gamma} = \min\Bigg\{&1 - \frac{\sum_{(r,c)\in\text{path}} n_{r,c}}{N + \left(\sum_{\forall(r,c)} n_{r,c}|r-c|^{\gamma}\right)^{1/\gamma}} \\ &+ \beta_1 \sum_{(r,c)\in\text{path}} n_{r,c}|r-c|^{\gamma} + \beta_2 N_{\text{disc\_pos}}\Bigg\}\end{aligned} \tag{8}$$

Now the minimization is performed over all possible paths from (1,1) to (K, K). Since $N_{\text{disc\_pos}}$ is a non-negative integer, setting $\beta_2 \geq 1$ will revert to the initial OC_β^γ. Note that $\text{OC}_{0;0}^1 = \frac{\text{MAE}}{1+\text{MAE}}$ is just a normalized version of MAE.

Nevertheless, we will not explore further this generalized index and all the following discussion will be based on the formulation (7).

3.2. *Single sample-size*

A key distinction between measures such as MAE (or MER or MSE) and Kendall's τ_b (or Spearman's rank correlation coefficient R_s or r_{int}) is that the latter cannot be applied to assess the performance in a single object. By working with pairs of observations, τ_b is not applicable to a single observation.

Although OC_β^γ integrates a ranking-based component, it is straightforwardly applied to a single example evaluation. Assume that the true and predicted classes of the observation correspond to the rth row and the cth column in the CM, respectively. Setting in Eq. (7), $N = 1$, $n_{r,c} = 1$, $n_{r',c'} = 0$ if $r', c' \neq r, c$, then OC_β^γ equals

$$\text{OC}_\beta^\gamma = \min\left(1; \quad 1 - \frac{1}{1 + |r - c|} + \beta|r - c|\right)$$

which increases monotonously from 0 to 1 when the distance of the example to the main diagonal increases from 0 to infinity. Figure 4 illustrates this evolution for different values of β. Note that, in this setting, for $\beta = 0.5$, OC already equals the MER.

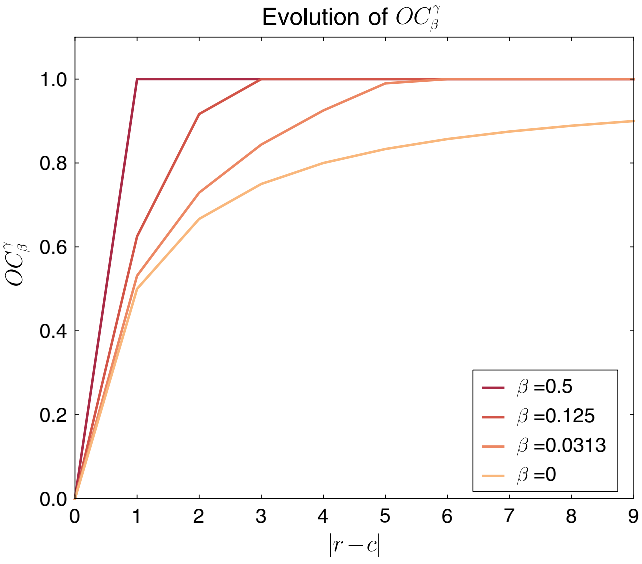


Fig. 4. Evolution of OC_β^γ for a single example evaluation.

3.3. *Properties of* $\mathbf{OC}_{\beta}^{\gamma}$

Let $\mathbf{a}, \mathbf{b}, \mathbf{c}$ be vectors used to construct CMs. It is easily observed from the definition that for $\beta > 0$, $\gamma > 1$ $\mathrm{OC}_{\beta}^{\gamma}(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$.

Since the cost given by (7) of any consistent path is always non-negative, $\mathrm{OC}_{\beta}^{\gamma}$ is always non-negative; since the cost of the path through the main diagonal is always not superior to 1, $\mathrm{OC}_{\beta}^{\gamma} \leq 1$.

It should be clear that the transposition of the CM does not change the value of $\mathrm{OC}_{\beta}^{\gamma}$ and therefore $\mathrm{OC}_{\beta}^{\gamma}$ is symmetric with respect to the role of the vectors involved in the construction of the CM: $\mathrm{OC}_{\beta}^{\gamma}(\mathbf{a}, \mathbf{b}) = \mathrm{OC}_{\beta}^{\gamma}(\mathbf{b}, \mathbf{a})$.

These conditions express intuitive notions about the expected properties for a classification performance index. It is also possible to establish that, for sufficiently high values of β, the triangular inequality is also satisfied, meaning that for certain values of β $\mathrm{OC}_{\beta}^{\gamma}$ is a metric. See Appendix A for further details.

3.4. *Computational remarks*

Noting from Eq. (7) that there is a cost $w_{r,c}$ corresponding to each vertex (entry in the matrix) of the graph given by

$$w_{r,c} = -\frac{n_{r,c}}{N + \left(\sum_{\forall(r,c)} n_{r,c}|r-c|^{\gamma}\right)^{1/\gamma}} + \beta n_{r,c}|r-c|^{\gamma}$$

the optimal consistent path can be found using dynamic programming. The first step is to traverse the matrix from the first entry to the last entry and compute the cumulative minimum weight W for all possible connected consistent paths for each entry (r, c):

$$W_{r,c} = w_{r,c} + \min\{W_{r-1,c-1}, W_{r-1,c}, W_{r,c-1}\}$$

with the adequate initialization ($W_{1,1} = 1 + w_{1,1}$) and the adequate attention for the entries in the first row and column. At the end of this process, the value $W_{K,K}$ will equal $\mathrm{OC}_{\beta}^{\gamma}$. The computational complexity of this process is $O(K^2)$.

For typical values of N and K, the overall complexity will be dominated by the cost of constructing the confusion matrix (N). This is also the complexity of MAE and MSE. Note also that the complexity of τ_b and r_{int} is not inferior to the complexity of OC.

4. Experimental Study

In this section we evaluate the behavior of the different coefficients in some additional cases, where it is possible to define a reasonable reference behavior. Typically, in the Minkowski distance, γ is rarely used for values other than 1, 2, and infinity. Since the overall conclusions do not differ for different γ values, we only present the experimental study for $\gamma = 1$. Simultaneously, the β values tested in this study are a percentage of the maximum possible value for the penalization term,

$N(K-1)^\gamma$. Since the choice for β is likely to be application dependent, balancing the tradeoff between the ranking and absolute classification, we present the results for two values of β, in the low and high range of the interval: $\beta_1 = \frac{0.25}{N(K-1)^\gamma}$ and $\beta_1 = \frac{0.75}{N(K-1)^\gamma}$.

4.1. *Tridiagonal matrices*

Consider CMs that are tridiagonal, taking the form

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & 1 & 0 & \cdots & 0 \\ & \vdots & & \ddots & & \vdots & \\ 0 & \cdots & 0 & 0 & 1 & 1 & 1 \\ 0 & \cdots & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Figure 5 plots the values of the coefficients for different number of classes. As the figure suggests and is analytically possible to conclude, r_{int}, R_s and τ_b all converge to 1 (perfect performance) as $K \to \infty$. In opposition, MER, MAE converge to 2/3 and OC_β^1 converges to 0.6. Our subjective evaluation of the performance of a classification result corresponding to a tridiagonal matrix would hardly correspond to the perfect performance. The r_{int}, R_s and τ_b coefficients seem therefore to present an unintuitive behavior. It is also interesting to discuss if the performance should improve with the increase of K. Subjectively, one may argue that with the increase of K, errors to the

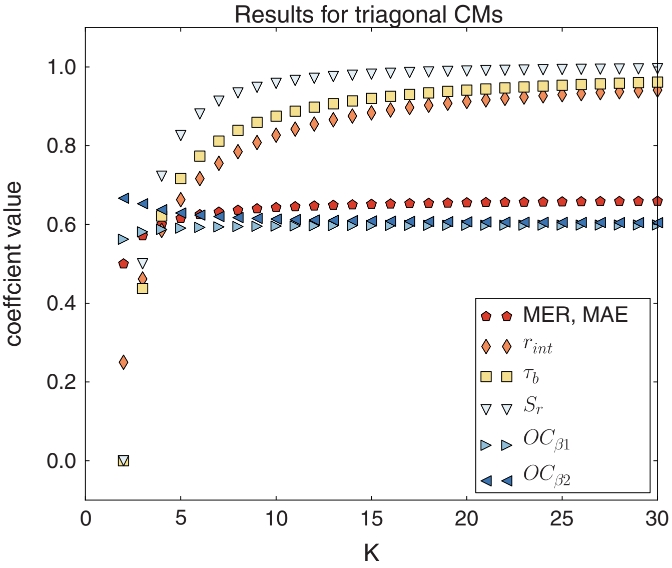


Fig. 5. Results for tridiagonal CMs, with $\beta_1 = \frac{0.25}{N(K-1)}$ and $\beta_2 = \frac{0.75}{N(K-1)}$.

sub- and super-diagonals of the CM become less significant and the performance should improve. Under this assumption, $\mathrm{OC}^1_{\frac{0.75}{N(K-1)}}$ presents the desired behavior.

4.2. *Dispersed examples*

To select the following examples, we randomly generated pairs of CMs and analyzed those where the relative performance as measure by OC_β did not agree with some of the other coefficients. Then, we tried to subjectively criticize the results.

A first pair of CMs is

$$\mathrm{CM}_1 = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 2 & 1 & 2 \end{bmatrix} \quad \mathrm{CM}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 2 & 2 & 1 \end{bmatrix}$$

The values for the coefficients we have been considering are provided in Table 2. All coefficients, except R_s and τ_b, seem to be in agreement with the expected conclusion that the performance corresponding to CM_2 is better than the performance corresponding to CM_1.

Consider now the pair of CMs

$$\mathrm{CM}_3 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 3 & 2 & 0 \end{bmatrix} \quad \mathrm{CM}_4 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

The values for the coefficients we have been considering are provided in Table 3. Now all coefficients, with the exception of r_{int}, seem to be in agreement with the expected conclusion that the performance corresponding to CM_4 is better than the performance corresponding to CM_3.

In a third example, consider the following CMs

$$\mathrm{CM}_5 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathrm{CM}_6 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Table 2. Results for CM_1 and CM_2, with $\beta_1 = \frac{0.25}{N(K-1)}$ and $\beta_2 = \frac{0.75}{N(K-1)}$.

CM	MER	MAE	R_s	τ_b	r_{int}	$\mathrm{OC}^1_{\beta_1}$	$\mathrm{OC}^1_{\beta_2}$
CM_1	0.50	0.80	0.20	0.19	0.39	0.63	0.69
CM_2	0.40	0.60	0.10	0.11	0.45	0.53	0.58

Table 3. Results for CM_3 and CM_4, with $\beta_1 = \frac{0.25}{N(K-1)}$ and $\beta_2 = \frac{0.75}{N(K-1)}$.

CM	MER	MAE	R_s	τ_b	r_{int}	$\mathrm{OC}^1_{\beta_1}$	$\mathrm{OC}^1_{\beta_2}$
CM_3	0.86	1.43	−0.26	−0.254	0.34	0.79	0.93
CM_4	0.57	0.85	−0.25	−0.250	0.08	0.71	0.75

Table 4. Results for CM_5 and CM_6, with $\beta_1 = \frac{0.25}{N(K-1)}$ and $\beta_2 = \frac{0.75}{N(K-1)}$.

CM	MER	MAE	R_s	τ_b	r_{int}	$\mathrm{OC}^1_{\beta_1}$	$\mathrm{OC}^1_{\beta_2}$
CM_5	0.86	1.00	0.89	0.84	0.81	0.58	0.75
CM_6	0.71	1.00	−0.29	−0.26	0.06	0.74	0.79

and the values in Table 4. This time MER and MAE were unable to capture the degradation of performance from CM_5 to CM_6. Note that CM_6 corresponds to an almost random classifier.

4.3. *Evaluation of real classifiers*

Following Ref. 9, we generated a synthetic dataset composed by 400 example points $\mathbf{x} = [x_1\ x_2]^t$ in the unit square $[0,1] \times [0,1] \subset \mathbb{R}^2$ according to a uniform distribution. Then, we assigned to each example $\mathbf{x}$ a class $y \in \{1,\ldots,5\}$ corresponding to

$$(b_0, b_1, b_2, b_3, b_4, b_5) = (-\infty; -2; -0.5; 0.25; 1; +\infty)$$

$$y = \min_{r\in\{1,2,3,4,5\}} \left\{ r : b_{r-1} < 10\left(\prod_{i=1}^{2} \mathbf{x}_i - 0.5\right) + \varepsilon < b_r \right\} \quad \varepsilon \sim \mathcal{N}(0, 0.125^2)$$

and represented in Fig. 6.

We compared the performance of three classifiers: the recently proposed data replication method,[4] instantiated both in Support Vector Machines (oSVM) and Neural Networks (oNN) and the method by Frank and Hall.[6] For completeness, we will briefly describe these learning techniques.

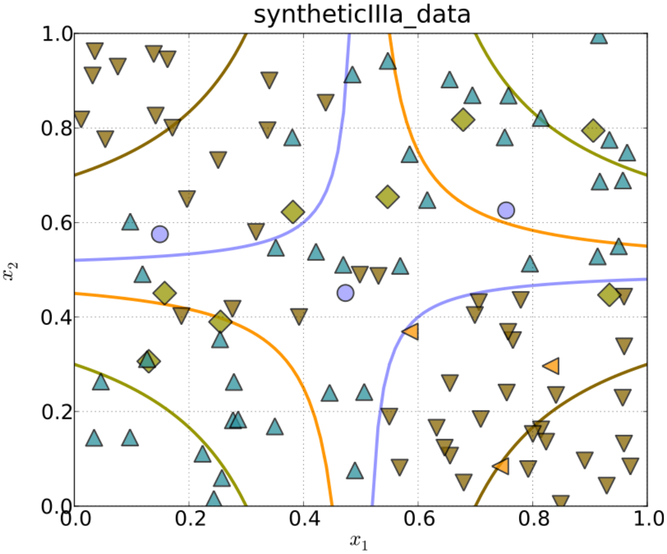


Fig. 6. Sample of 100 examples from synthetic dataset ($K = 5$).

The data replication method for ordinal data can be framed under the single binary classifier reduction (SBC), an approach for solving multiclass problems via binary classification relying on a single, standard binary classifier. SBC reductions can be obtained by embedding the original problem in a higher-dimensional space consisting of the original features, as well as one or more other features determined by fixed vectors, designated here as *extension features*. This embedding is implemented by replicating the training set points so that a copy of the original point is concatenated with each of the extension features' vectors. The binary labels of the replicated points are set to maintain a particular structure in the extended space. This construction results in an instance of an artificial binary problem, which is fed to a binary learning algorithm that outputs a single binary classifier. To classify a new point, the point is replicated and extended similarly and the resulting replicas are fed to the binary classifier, which generates a number of signals, one for each replica. This method can be instantiated in two important machine learning algorithms: support vector machines and neural networks. For more details, the reader should consult Ref. 4.

Frank and Hall in Ref. 6 proposed to use $(K-1)$ standard binary classifiers to address the K-class ordinal data problem. Towards that end, the training of the ith classifier is performed by converting the ordinal dataset with classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$ into a binary dataset, discriminating $\mathcal{C}_1, \ldots, \mathcal{C}_i$ against $\mathcal{C}_{i+1}, \ldots, \mathcal{C}_K$. The ith classifier represents the test $\mathcal{C}_x > \mathcal{C}_i$. To predict the class value of an unseen instance, the $K-1$ binary outputs are combined to produce a single estimation.

Using the aforementioned techniques, the dataset was split in 40% for training ($\mathcal{D}$) and 60% for testing ($\mathcal{D}^*$). Algorithm 1 illustrates the experimental procedure. The splitting of the data was repeated 50 times in order to obtain more stable results for performance estimation. In lines 6 and 16 of Algorithm 1 one can use any of the metrics discussed in this manuscript in order to obtain the best parameterization of the model or estimate the final performance.

In the results of Table 5, CM_{10} represents the results for oSVM, CM_{11} the result for oNN and CM_{12} the performance for Frank and Hall. The CMs are as follows:

$$\mathrm{CM}_{10} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 50 & 7 & 0 & 0 \\ 0 & 2 & 94 & 2 & 0 \\ 0 & 0 & 11 & 39 & 0 \\ 0 & 0 & 0 & 5 & 30 \end{bmatrix} \quad \mathrm{CM}_{11} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 45 & 12 & 0 \\ 0 & 0 & 2 & 87 & 9 \\ 0 & 0 & 0 & 6 & 44 \\ 0 & 0 & 0 & 0 & 35 \end{bmatrix}$$

$$\mathrm{CM}_{12} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 50 & 7 & 0 & 0 \\ 0 & 2 & 94 & 2 & 0 \\ 0 & 0 & 21 & 29 & 0 \\ 0 & 0 & 0 & 29 & 6 \end{bmatrix}$$

```
Data: $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ the training dataset and $D^* = \{\mathcal{X}^*, \mathcal{Y}^*\}$ the testing set.
Result: $\mathcal{M}$, trained model, accuracy accuracy result for $\mathcal{D}^*$ and respective
        CM.
Best_Accuracy $\leftarrow 0$;
Partition training data $\mathcal{D}$ in five equal subsets so that
$\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\} = \{(\mathcal{X}^1, \mathcal{Y}^1) \cup \ldots \cup (\mathcal{X}^5, \mathcal{Y}^5)\}$
forall the parameterization values $p$ do
    forall the $fold = 1$ to 5 do
        $\mathcal{M} \leftarrow$ Train_Model($\mathcal{X}^i, \mathcal{Y}^i, p$)   where $i = \{1, \ldots, 5\} \backslash fold$
        $\mathcal{Y}_1 \leftarrow$ Test_Model($\mathcal{M}, \mathcal{X}^{fold}$);
        accuracy$^{fold}$ $\leftarrow$ assess performance according a given measure, m,
        ($\mathcal{Y}_1, \mathcal{Y}^{fold}$);
    end
    $\overline{\text{accuracy}} \leftarrow 1/5 \sum_{i=1}^{5}$ accuracy$^i$;
    if $\overline{\text{accuracy}}$ > Best_Accuracy then
        Best_Accuracy $\leftarrow \overline{\text{accuracy}}$ ;
        Best_Parameterization $\leftarrow p$;
    end
end
$\mathcal{M} \leftarrow$ Train_Model($\mathcal{X}, \mathcal{Y}$, Best_Parameterization);
($\mathcal{Y}_1, CM$) $\leftarrow$ Test_Model($\mathcal{M}, \mathcal{X}^*$);
accuracy $\leftarrow$ assess performance according a given measure, m, ($\mathcal{Y}_1, \mathcal{Y}^*$);
```

Algorithm 1. Experimental procedure to design the models. This procedure was repeated 50 times in order to obtain more stable results for performance estimation.

A subjective analysis of the CMs places CM_{10} as the best result and CM_{11} in the bottom. Although all indices capture this relative performance, R_s, τ_b and r_{int} almost do not differentiate CM_{11} from CM_{12}. The Ordinal Classification Index, on the other hand, portrays a significant difference in performance, in spite of also incorporating a ranking term.

4.4. *Experiments with real datasets*

To further evaluate the impact of using OCI, we performed the following experiments with sets of real ordinal data, testing our method on the SWD, LEV, ESL,

Table 5. Results for CM_{10}, CM_{11} and CM_{12}, with $\beta_1 = \frac{0.25}{N(K-1)}$ and $\beta_2 = \frac{0.75}{N(K-1)}$.

CM	MER	MAE	R_s	τ_b	r_{int}	$\text{OC}^1_{\beta_1}$	$\text{OC}^1_{\beta_2}$
CM_{10}	0.11	0.11	0.93	0.91	0.91	0.12	0.13
CM_{11}	0.82	0.91	0.89	0.85	0.84	0.55	0.66
CM_{12}	0.25	0.25	0.90	0.86	0.86	0.23	0.26

Balance and BCCT datasets. The first dataset, SWD, contains real-world assessments of qualified social workers regarding the risk facing children if they stayed with their families at home and is composed by ten features and four classes. LEV dataset contains examples of anonymous lecturer evaluations, taken at the end of MBA courses and is composed by four features and five classes. These datasets contain 1000 examples each.

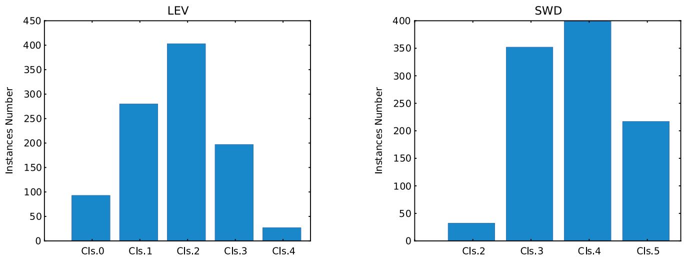
Another dataset which we worked on was the ESL dataset containing 488 profiles of applicants for certain industrial jobs. Features are based on psychometric test results and interviews with the candidates performed by expert psychologists. The class assigned to each applicant was an overall score corresponding to the degree of fitness for the type of job.

Balance dataset available on UCI machine learning repository was also experimented. Created to model psychological experimental results, each example is labeled as having a balance scale tip to the right, left or balanced. Features encompass on left and right weights, and distances.

The last dataset encompasses on 1144 observations taken from previous works[3] and expresses the aesthetic evaluation of Breast Cancer Conservative Treatment (BCCT). For each patient submitted to BCCT, 30 measurements were recorded, capturing visible skin alterations or changes in breast volume or shape. The aesthetic outcome of the treatment for each and every patient was classified in one of the four categories: Excellent, Good, Fair and Poor. In Fig. 7 is depicted the class frequency distribution for each dataset.

To assess the merit of OCI in an ordinal data classification setting, we trained three different classifiers on the five mentioned datasets:

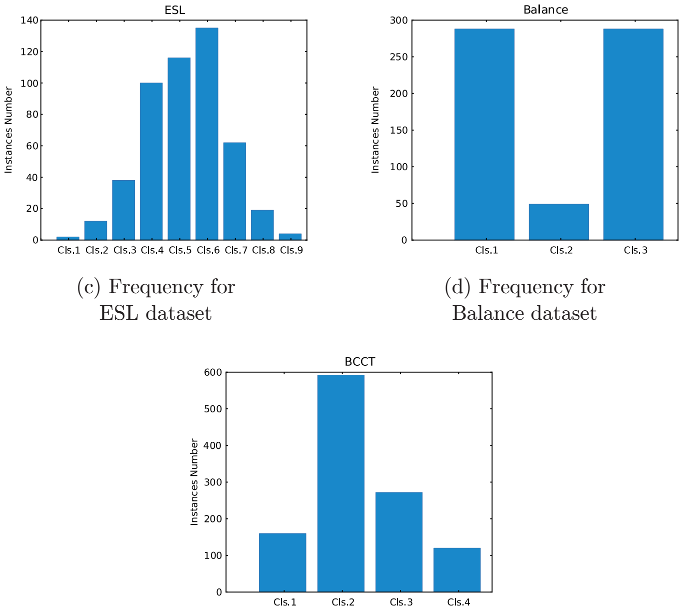
- A conventional multiclass classifier, based on the one-against-one rationale. The baseline binary classifier was the binary SVM, as deployed in libSVM.[5]
- The multiclass classifier adapted for ordinal data based on the proposal by Frank and Hall, as described previously. The baseline binary classifier was again the



(a) Frequency for LEV dataset

(b) Frequency for SWD dataset

Fig. 7. Real datasets frequency values.



(c) Frequency for ESL dataset

(d) Frequency for Balance dataset

(e) Frequency for BCCT dataset

Fig. 7. (*Continued*)

binary SVM, as deployed in libSVM. Previous works have shown the advantage of this method over conventional approaches.[6,9]

- The data replication method, instantiated in SVMs (oSVM), as also described before. Previous works have shown the advantage of this method over both conventional approaches and the Frank and Hall method.[3,15,16]

Once again the experimental study followed the setting illustrated in Algorithm 1. The datasets were split in 40% for training and 60% for testing; the optimization of the parameters using cross-validation over the training set was based on the OCI metric; the final assessment of the performance of the models in the test set was done again using OCI. A linear kernel was used in all learning schemes. The results are presented in Table 6.

Table 6. Performance average (std. dev.) results for the five datasets using the OCI measure.

Dataset	oSVM	Frank and Hall	Conventional
SWD	0.49 (0.02)	0.47 (0.01)	0.49 (0.02)
LEV	0.44 (0.02)	0.46 (0.02)	0.47 (0.02)
ESL	0.36 (0.00)	0.36 (0.01)	0.36 (0.01)
Balance	0.13 (0.01)	0.13 (0.01)	0.14 (0.02)
BCCT	0.39 (0.01)	0.39 (0.01)	0.40 (0.01)

A first main assertion is that OCI correctly captures the superiority of both algorithms specific to ordinal data over the conventional method. The learning and the assessment with OCI are in accordance with the expected relative performance. The relative merit of oSVM and Frank and Hall method is not that strong, with a potentially slight advantage of oSVM, never losing for Frank and Hall method, both in average and in variance. It is also important to notice that oSVM produces simpler models than Frank and Hall method, since all boundaries share the same direction (the boundaries) are parallel. Likewise, Frank and Hall method produces simpler and more robust classifiers than the one-against-one generic model implemented in libSVM.

5. Conclusion

We have proposed the use of a metric defined directly on the CM to evaluate the performance in ordinal data classification. The metric chooses the non-discordant pairs of observations that minimize the cost of a global optimization procedure on the CM, minimizing deviation of the pairs to the main diagonal while maximizing the benefit. The adoption of this measure thus guarantees fair comparison among competing systems, and more correct optimization procedures for classifiers.

Arguing in favor of a new metric against current ones is a difficult task, almost requiring a meta-metric to assess the performance of metrics. To overcome this difficulty we started by trying to motivate the interest of the proposed metric with intuitive settings and completed with the application in real datasets.

A new metric can be used not only to compare classifiers but also to design better classifiers. The usage in the design of classifiers can be in two different directions. A first use is "externally" to the classifier, using the metric to select the best parameterization of the classifier; in this paper we have used the metric for optimizing the parameters of the models using cross-validation. A second possibility is to embed the new metric in the classifier design, adapting the internal objective function of the classifier, replacing loss functions based on standard measures by a loss function based on the proposed measure. For instance, the standard loss function of a neural network based on the square of the error or on cross-entropy could be replaced by an error evaluated by OCI. This will be pursued in future research.

Acknowledgments

This work was partially funded by Fundação para a Ciência e a Tecnologia (FCT) — Portugal through project PTDC/EIA/64914/2006. The authors are grateful to the reviewers for their many stimulating and thoughtful comments.

Appendix A. Triangular Inequality

For sufficiently high values of $\beta (\beta \geq \frac{1}{N+1})$ the optimal path is always over the main diagonal and the ordinal classification index simplifies to $1 - \frac{\sum_{(r,c) \in \text{main diagonal}} n_{r,c}}{N + \left(\sum_{\forall (r,c)} n_{r,c} |r-c|^{\gamma}\right)^{1/\gamma}} = \frac{M+H}{M+N} = \frac{M}{M+N} + \frac{H}{M+N}$, where H and M are the Hamming and Minkowski distances, respectively. This is easily seen to be a metric:

- The positive definiteness and symmetry have already been established in the main body of the article;
- Knowing that if d_1 and d_2 are metrics and $d_1(\mathbf{a}, \mathbf{b}) \leq d_2(\mathbf{a}, \mathbf{b}), \ \forall \mathbf{a}, \mathbf{b}$, then

(1) $\frac{d_2}{1+d_2}$ is a metric;
(2) $\frac{d_1}{1+d_2} \leq \frac{d_2}{1+d_2}$ is a metric;
(3) $d_1 + d_2$ is a metric;

It just remains to prove that for $\beta \geq 1/(N+1)$ the optimal path is indeed the main diagonal. Let p be a consistent path and b_1 be the part of benefit of the path on the main diagonal and $b_2 > 0$ the part of benefit of the path not in the main diagonal. If $\beta \geq \frac{1}{N+1}$ then the following is true for the cost C of the path:

$$\begin{aligned} C &= 1 - \frac{b_1 + b_2}{N+M} + \beta \sum_{(r,c) \in \text{path}} n_{r,c} |r-c|^{\gamma} \\ &\geq 1 - \frac{b_1}{N+M} - \frac{b_2}{N+M} + \frac{1}{N+1} \sum_{(r,c) \in \text{path}} n_{r,c} |r-c|^{\gamma} \\ &\geq 1 - \frac{b_1}{N+M} - \frac{b_2}{N+M} + \frac{b_2}{N+1} \geq 1 - \frac{b_1}{N+M} \end{aligned}$$

This last value is clearly not inferior to the cost of the path over the main diagonal.

To finalize, it is easy to conclude that for small values of β, $\mathrm{OC}_{\beta}^{\gamma}$ is not a metric. Consider the vectors ($K = 2$)

$$\mathbf{a} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ 2 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 2 \\ 1 \end{bmatrix}$$

The corresponding confusion matrices are

$$\mathrm{CM}(\mathbf{a}, \mathbf{b}) = \begin{bmatrix} N-1 & 0 \\ 1 & 0 \end{bmatrix} \quad \mathrm{CM}(\mathbf{b}, \mathbf{c}) = \begin{bmatrix} N-1 & 1 \\ 0 & 0 \end{bmatrix} \quad \mathrm{CM}(\mathbf{a}, \mathbf{c}) = \begin{bmatrix} N-2 & 1 \\ 1 & 0 \end{bmatrix}$$

```
% input: confusion matrix and number of classes
% size(cMatrix) must be [K K]
function oc=OrdinalClassificationIndex(cMatrix, K)
  N = sum(cMatrix(:));
  ggamma = 1;
  bbeta  = 0.75/(N*(K-1)^ggamma);

  helperM2 = zeros(K,K);
  for r=1:K
    for c=1:K
      helperM2(r,c) = cMatrix(r,c) * ((abs(r-c))^ggamma);
    end
  end
  TotalDispersion=(sum(helperM2(:))^(1/ggamma));
  helperM1        =cMatrix/(TotalDispersion+N);

  errMatrix(1,1) = 1 - helperM1(1,1) + bbeta*helperM2(1,1);
  for r =2:K
    c=1;
    errMatrix(r,c) = errMatrix(r-1, c) - helperM1(r,c) + ...
        bbeta*helperM2(r,c);
  end
  for c =2:K
    r=1;
    errMatrix(r,c) = errMatrix(r,c-1) - helperM1(r,c) + ...
        bbeta*helperM2(r,c);
  end

  for c=2:K
    for r=2:K
      costup      = errMatrix(r-1, c);
      costleft    = errMatrix(r, c-1);
      lefttopcost = errMatrix(r-1, c-1);
      [aux,idx]   = min([costup costleft lefttopcost]);
      errMatrix(r,c) = aux - helperM1(r,c) + bbeta*helperM2(r,c);
    end
  end
  oc = errMatrix(end,end);
  return
```

Listing 1. Ordinal classification index computation.

It is easy to confirm that for $\beta < \frac{N-1}{(N+1)(N+2)}$ we have $\mathrm{OC}_\beta^\gamma(\mathbf{a},\mathbf{b}) + \mathrm{OC}_\beta^\gamma(\mathbf{b},\mathbf{c}) < \mathrm{OC}_\beta^\gamma(\mathbf{a},\mathbf{c})$ and therefore OC_β^γ does not obey the triangular inequality.

Appendix B. Source Code Listing

For reference, Listing 1 is presented for a Matlab implementation of OC_β^γ.

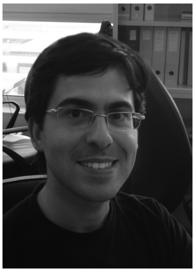
References

1. S. Baccianella, A. Esuli and F. Sebastiani, Evaluation measures for ordinal regression, *Proc. Ninth Int. Conf. Intelligent Systems Design and Applications*, 2009, pp. 283–287.

2. A. Ben-David, A lot of randomness is hiding in accuracy, *Eng. Appl. AI* **20**(7) (2007) 875–885.
3. J. S. Cardoso and M. J. Cardoso, Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment, *Artif. Intell. Med.* **40** (2007) 115–126.
4. J. S. Cardoso and J. F. Pinto da Costa, Learning to classify ordinal data: The data replication method, *J. Mach. Learn. Res.* **8** (2007) 1393–1429.
5. C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001, Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
6. E. Frank and M. Hall, A simple approach to ordinal classification, *Proc. 12th European Conf. Machine Learning* (Springer-Verlag, 2001), pp. 145–156.
7. L. Gaudette and N. Japkowicz, Evaluation methods for ordinal classification, in *Proc. 2nd Canadian Conf. Artificial Intelligence*, eds. Y. Gao and N. Japkowicz, Lecture Notes in Computer Science (Springer, 2009), pp. 207–210.
8. T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, Monographs on Statistics and Applied Probability, Vol. 43 (1990), pp. 297–318.
9. R. Herbrich, T. Graepel and K. Obermayer, Support vector learning for ordinal regression, *Ninth Int. Conf. Artificial Neural Networks ICANN* **1** (1999) 97–102.
10. L. Jiang, D. Wang, H. Zhang, Z. Cai and B. Huang, Using instance cloning to improve naive Bayes for ranking, *Int. J. Patt. Recogn. Artif. Intell.* **6** (2008) 1121–1140.
11. M. Kendall, A new measure of rank correlation, *Biometrika* **30** (1938) 81–89.
12. S. Kotsiantis and D. Kanellopoulos, Cascade generalisation for ordinal problems, *Int. J. Artif. Intell. Soft Comput.* **2** (2010) 46–57.
13. J. W. T. Lee and D.-Z. Liu, Induction of ordinal decision trees, *Proc. Int. Conf. Machine Learning and Cybernetics* **4** (2002) 2220–2224.
14. P. McCullagh, Regression models for ordinal data, *J. Roy. Stat. Soc. Series B* **42** (1980) 109–142.
15. J. F. Pinto da Costa, H. Alonso and J. S. Cardoso, The unimodal model for the classification of ordinal data, *Neural Networks* **21** (2008) 78–91.
16. J. Pinto da Costa, R. Sousa and J. S. Cardoso, An all-at-once unimodal SVM approach for ordinal classification, *Proc. Ninth Int. Conf. Machine Learning and Applications (ICMLA 2010)*, 2010.
17. W. Press, B. Flannery, S. Teukolsky and W. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 2002).
18. V. C. Raykar, R. Duraiswami and B. Krishnapuram, A fast algorithm for learning a ranking function from large-scale data sets, *IEEE Trans. Patt. Anal. Mach. Intell.* **30**(7) (2008) 1158–1170.
19. A. Shashua and A. Levin, Ranking with large margin principle: Two approaches, *Adv. Neural Inform. Process. Syst.* **15**, eds. Thrun and K. Obermayer (MIT Press, 2003), pp. 937–944.
20. C. Spearman, The proof and measurement of association between two things, *American J. Psychol.* **15** (1904) 72–101.
21. G. Tutz, Generalized semiparametrically structured ordinal models, *Biometrics* **59** (2003) 263–273.
22. S. Vanbelle and A. Albert, A note on the linearly weighted kappa coefficient for ordinal scales, *Statist. Methodol.* **6**(2) (2009) 157–163.
23. W. Waegeman, *Learning to Rank: a ROC-Based Graph-Theoretic Approach*, Ph.D. thesis, Universiteit Gent, 2009.
24. W. Waegeman, B. De Baets and L. Boullart, A comparison of different ROC measures for ordinal regression, *Proc. CML 2006 Workshop on ROC Analysis in Machine Learning* (2006).

Jaime S. Cardoso received his B.Sc. in electrical and computer engineering from Faculdade de Engenharia, Universidade do Porto, in 1999. He also holds a Masters in engineering mathematics from Faculdade de Ciências, Universidade do Porto, 2005 and holds a Ph.D. in computer vision from Universidade do Porto in 2006.

Since 2006, Cardoso has been an assistant professor at FEUP, where he teaches machine learning and computer vision related courses. He is currently a project leader at INESC Porto. His main research interests include visual pattern recognition and machine learning. He is also Senior Member of IEEE and co-founder of ClusterMedia Labs, an IT company developing automatic solutions for semantic audio-visual analysis.



Ricardo Sousa holds a B.Sc. in computer science and M.Sc. in mathematical engineering from the Universidade do Porto in 2007 and 2008, respectively. Currently, he is working towards his Ph.D. degree in machine learning at the Universidade do Porto. He was an intern student at Federal University of Ceará, Brazil in 2011. Since 2007, he has been involved in the collaboration of several research projects at INESC Porto.

His research interests include machine learning and image processing.