Metrics to Guide a Multi-Objective Evolutionary Algorithm for Ordinal Classification

M. Cruz-Ramírez*, C. Hervás-Martínez, J. Sánchez-Monedero, P.A. Gutiérrez

Department of Computer Science and Numerical Analysis. University of Córdoba, Spain

Abstract

Ordinal classification or ordinal regression are classification problems in which the labels have an ordered arrangement between them. Due to this order, alternative performance evaluation metrics are need to be used in order to consider the magnitude of errors. This paper presents an study of the use of a multi-objective optimization approach in the context of ordinal classification. We contribute a study of ordinal classification performance metrics, and propose a new performance metric, the Maximum Mean Absolute Error ($MMAE$). $MMAE$ considers per-class distribution of patterns and the magnitude of the errors, both issues being crucial for ordinal regression problems. In addition we empirically show that some of the performance metrics are competitive objectives, which justifies the use of multi-objective optimization strategies. In our case, a multi-objective evolutionary algorithm optimizes a artificial neural network ordinal model with different pairs of metrics combinations, and we conclude that the pair of the Mean Absolute Error (MAE) and the proposed $MMAE$ is the most favorable. A study of the relationship between the metrics of this proposal is performed, and the graphical representation in the 2 dimensional space where the search of the evolutionary algorithm takes place is analyzed. The results obtained show a good classification performance, opening new lines of research in the evaluation and model selection of ordinal classifiers.

Keywords: mean absolute error, multi-objective evolutionary algorithm, ordinal measures, ordinal classification, ordinal regression, proportional odds model

1. Introduction

Ordinal classification or ordinal regression is a supervised learning problem of predicting categories that have an ordered arrangement. Although classification and regression metric problems have been thoroughly investigated in the literature, the ordinal regression problems have not received as much attention as nominal (binary or multiclass) classification. For example, people can be classified by considering whether they are high, medium, or low on some attribute or in a set of categories varying from strong agreement to strong disagreement with respect to some attitude item. Hodge and Treiman [1], to analyze social class identification, scored responses as follows: "Respondents identifying with the lower, working, middle, upper middle, and upper class were assigned the scores 1, 2, 3, 4, and 5, respectively". Though sequential numbers may be assigned to such categories, the numbers assigned serve only to identify the ordering of the categories. In contrast to regression metric problems, these ranks are finite types and the metric distances between the ranks are not defined, in general; in contrast to classification problems, these ranks are also different from the labels of multiple classes due to the existence of the ordering information [2].

In the previous example, it is straightforward to think that predicting class *lower* when the real class is *upper middle* should be considered as a more severe error than the one associated to a *working* prediction. Thereby, ordinal classification problems should be evaluated with specific metrics. On a first consideration, various measures of ordinal association and product-moment correlation and regression seem to rely on very different

*Corresponding author at: Department of Computer Science and Numerical Analysis, University of Córdoba, Rabanales Campus, Albert Einstein Building 3rd Floor, 14071 Córdoba, Spain. Tel.: +34 957 218 349; Fax: +34 957 218 630.

E-mail addresses: {mcruz,chervas,jsanchezm,pagutierrez}@uco.es

**This paper is a significant extension of the work "A Preliminary Study of Ordinal Metrics to Guide a Multi-Objective Evolutionary Algorithm" appearing in the 11th International Conference on Intelligent Systems Design and Applications (ISDA2011).

foundations. This is, the ordinal measures are developed from a) the notion of comparing pairs of cases, or b) the product-moment system, which is considered in terms of measures of individual cases.

If methodology a) is used, and there is an ordering of the categories but the absolute distances among them are unknown, an ordinal categorical variable is obtained. In that respect, in order to avoid the influence of the numbers chosen to represent the classes on the performance assessment, we should only look at the order relation between "true" and "predicted" class numbers. The use of Spearman's rank correlation coefficient, r_S, [3] and specially Kendall's τ_b [4] is a step forward in that direction. Moreover, other coefficients are frequently used to describe association between ordinal measures as Goodman and Kruskal's γ [5], and Somers's d [6].

If methodology b) (product-moment system) is used, the most common considered measures in machine learning are the Mean Absolute Error (here denoted as MAE) [7, 8], Root Mean Square Error ($RMSE$) [8], and Mean Zero-One Error (MZE, more frequently known as error rate) [8], being $MZE = 1 - CCR$, where CCR is the Correct Classification Rate. However, these three measures are not suitable when used to evaluate the performance of classifiers on ordinal unbalanced datasets [7]. The first contribution of this work is a newly proposed metric associated to an ordinal classifier that is the highest MAE value from MAEs measured independently for each class (Maximum MAE or $MMAE$). This metric evaluates the performance on the worst classified class. The second contribution of this work is the analysis of the state-of-the-art performance metrics. Finally, we empirically show that some of the metric pairs can be non-cooperative, and consequently justify the use of a multi-objective framework to address the classifier optimization problem.

Figure 1 presents a motivational example for the present work depicting three classifiers on a fourth class ordinal classification problem. This figure illustrates how different variations of decision thresholds can affect to classification performance specially influenced by patterns placed on the classes boundaries. More specifically, this example raise two issues that will be studied in the current work. Firstly, using a unique performance measure may be not enough to evaluate a classifier, specially in the field of ordinal regression. Second, some of the performance metrics can result on competitive objectives on a general optimization process since moving a threshold on a direction can produce an improvement in one metric, but a detrimental on a second one.

In the present paper, the aforementioned issues are studied under a multi-objective optimization approach. Multi-objective algorithms are algorithms that optimize simultaneously objectives that are non-cooperative. In many problems there are several conflicting objectives, such as execution speed or computational cost and kindness of the results. For example, in [9, 10] the authors try to obtain optimal results in the shortest time and at the lowest cost. In other problems, the execution speed is not the most important and what is relevant is achieving good results in different conflicting error functions.

In the field of Artificial Neural Networks (ANNs), classification performance and model simplicity are objectives that typically guide the training process of a Evolutionary Multi-Objective Algorithm (MOEA) [11], with the purpose of finding a trade-off between performance and model readability. Other works present the optimization of global performance versus worst classified class in a Pareto based algorithm [12] or also by simplifying both objectives as a weighted linear combination of the functions [13].

In ordinal classification, it is common to use several error functions when some of the classes have a number of patterns much lower than the others, i.e. ordinal imbalanced datasets. Because of this reason, we proposed the $MMAE$ metric measuring the performance in the worst classified class. One real world application where this problem can be found is in the extension of donor-recipient allocation in liver transplants [14], where the classifiers aims at predicting the survival of the organ (describing this survival in three different classes, class 1: lower than 15 days, class 2: between 15 days and 3 months and, class 3: higher than 3 months). The problem is that, in real cases, the number of patterns of class 1 is much lower than that of class 2 or 3. The hospital would be interested in classifiers able to correctly classify all classes equally, but the bad performance for class 1 can be hidden by the fact that the number of patterns of this class is very low (for example, a good MAE value can be obtained when class 1 is associated to a 5% of the patterns and the classifier never assigns a pattern to class 1). As can be seen, both objectives are conflicting (MAE and $MMAE$), because improving $MMAE$ usually involves worsening MAE and vice versa. In [15] another ordinal problem is solved from a multi-objective perspective, where six different objectives are considered, including MZE, MAE and four different formulations for the expected ranking accuracy. In this work, several different ordinal measures that could be combined in the context of ordinal regression are analyzed and combined in pairs for a MOEA.

The present work aims at identifying which pair of

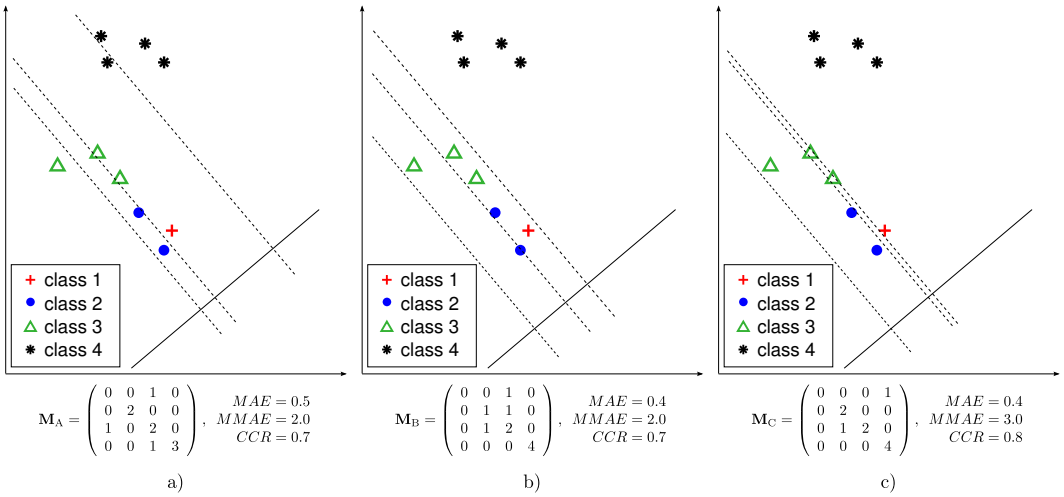

Figure 1: An example of three classifiers decision boundaries for a four class ordinal classification problem. Decision thresholds vary from right to left leading to three different situations regarding performance evaluation metrics. Classifiers on subfigure a) and b) have the same CCR and $MMAE$, whereas MAE varies and confusion matrices M_A and M_B are different. Similar comment applies when comparing b) and c) situations, but in this case MAE is kept constant while CCR and $MMAE$ vary. Finally, when comparing the classifiers a) and c), the CCR and MAE values improve and the value of $MMAE$ worsens.

ordinal classification performance metrics can be more suitable to guide a MOEA to obtain classifiers with a good performance (considering both the order of the miss-classification errors and the worst classified class errors). The most common ordinal classification performance metrics are reviewed, and some of them are selected to evaluate the performance of four nominal and ordinal classifiers, including also the proposed metric. Then, a correlation study is done between all the metrics in order to find the less correlated ones. We hypothesize that the more uncorrelated metrics are the more suitable for acting as optimization objectives for the MOEA (given that all of them highlight positive aspects of the classifiers). The selected metrics are grouped into different pairs that will be simultaneously optimized by the MOEA. The base classifier considered is an ANN based on the Proportional Odds Model (POM) [16] and it is evolved using a differential evolution MOEA [17, 18]. Finally, the generalization performance of the models obtained is studied with respect to the pair of metrics considered in the evolution. Because of their performance, the pair $MMAE$-MAE is taken special attention, deriving a relationship between this pair of metrics and studying their graphical representation.

This paper is a significant extension of a conference paper [19]. The new contributions are the following. A correlation analysis of the ordinal performance metrics is done, and the confusion matrices studied are changed and extended. In addition, the neural network model used has been replaced by another neural network based on the Proportional Odds Model and a local search procedure based on the iRprop$^+$ algorithm [20] has been included in the MOEA to optimize the new model. Finally, two additional ordinal methods have been compared in the experimental section, and new tables describing the experiments and statistical tests have been included to enforce the conclusions.

The rest of paper is organized as follows: Section 2 shows a revision and an experimental comparison of measures for ordinal classification; Section 3 details the ordinal ANN model based on the POM model; Section 4 describes the training method employed; Section 5 describes the experimental design and the results obtained, while conclusions and future research are outlined in Section 6.

2. Measures of association in ordinal classification

This section presents both nominal and ordinal classification performance metrics commonly used in the literature. An empirical evaluation of the correlation between them is done in order to select the most relevant ones.

Let's define an ordinal classification problem as a problem where the purpose is to learn a model able to predict class labels, $\mathcal{C} = \{C_1, C_2, \ldots, C_J\}$ containing J

Table 1: Confusion matrix.

		Predicted class					
		1	$\cdots$	k	$\cdots$	J	
Real class	1	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1J}	$n_{1\bullet}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	j	n_{j1}	$\cdots$	n_{jk}	$\cdots$	n_{jJ}	$n_{j\bullet}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	J	n_{J1}	$\cdots$	n_{Jk}	$\cdots$	n_{JJ}	$n_{J\bullet}$
		$n_{\bullet 1}$	$\cdots$	$n_{\bullet k}$	$\cdots$	$n_{\bullet J}$	n

labels, for unseen patterns after a training process. What makes the difference with nominal classification is that the label set has an order relation $C_1 \prec C_2 \prec \ldots \prec C_J$ imposed on it (the symbol $\prec$ denotes the ordering between different ranks). Let's suppose an ordinal classification problem with J classes and n patterns with a classifier g obtaining a $J \times J$ contingency or confusion matrix. Table 1 shows the confusion matrix, where n_{jk} represents the number of times the patterns are predicted by classifier g to be in class k when they really belong to class j, $n_{j\bullet}$ is the number of patterns belonging to class j, $n_{\bullet k}$ is the number of patterns predicted in class k and n is the total number of patterns.

Let $D = \{(\mathbf{x}_i, y_i),\ i = 1, \cdots, n\}$ be the set of training patterns, denote by $\{y_1, y_2, \ldots, y_n\}$ the set of labels of a given dataset, and let $\{y_1^*, y_2^*, \ldots, y_n^*\}$ be the predicted labels by the evaluated classifier g, where $y_i \in C$ and $y_i^* \in C$, and $C = \{C_1, C_2, \ldots, C_J\}$, $1 \leq i \leq n$. In general, the predictions of the classifiers will be categories but, for some metrics, these categories will be turned into integer values by using the function $O(y_i^*)$ which establishes the position on the ordinal scale of the predicted label, being $O(C_j) = j$ and if $y_i = C_j$ them $O(y_i) = j$, $1 \leq j \leq J$.

Many ordinal measures have been proposed to determine the efficiency of the classifier g, but not all pairs formed by these metrics might be valid to guide a MOEA. Before describing the ordinal metrics, accuracy and minimum sensitivity are also presented, since they have been proved to be non-cooperative objectives [12] and they are commonly used (especially the CCR) in classification problems:

- CCR: The Correct Classification Rate or accuracy is the percentage of correctly classified patterns:

$$CCR = \frac{1}{n} \sum_{j=1}^{J} n_{jj},$$

  where CCR values range from 0 to 1.

- MS: The Minimum Sensitivity is the lowest percentage of patterns correctly predicted as belonging to each class, with respect to the total number of examples in the corresponding class:

$$MS = \min\{S_j = \frac{n_{jj}}{n_{j\bullet}};\ j = 1, ..., J\},$$

  where S_j is the sensitivity of the *j-th* class and MS values range from 0 to 1.

On the other hand, there are other product-moment ordinal metrics specifically used in ordinal classification:

- MAE: The Mean Absolute Error is the average absolute deviation of the predicted class from the true class (i.e. average absolute deviation in number of categories of the ordinal scale) [7]:

$$MAE = \frac{1}{n} \sum_{j,k=1}^{J} |j - k| n_{jk} = \frac{1}{n} \sum_{i=1}^{n} e(\mathbf{x}_i),$$

  where $e(\mathbf{x}_i) = |O(y_i) - O(y_i^*)|$ is the distance between the true (y_i) and the predicted (y_i^*) ranks, and $O(C_j) = j$ is the position of a label in the ordinal rank. Then, MAE values range from 0 to $J - 1$.

- $AMAE$: The Average MAE is the mean of the MAE classification errors across classes and was proposed by Baccianella et al. [7] to mitigate the effect of imbalanced class distributions. Let MAE_j be the MAE for a given *j-th* class:

$$MAE_j = \frac{1}{n_{j\bullet}} \sum_{k=1}^{J} |j - k| n_{jk} = \frac{1}{n_{j\bullet}} \sum_{i=1}^{n_{j\bullet}} e_j(\mathbf{x}_i),\ 1 \leq j \leq J,$$

  in such a way that:

$$MAE = \frac{1}{n} \sum_{j=1}^{J} n_{j\bullet} MAE_j.$$

  $AMAE$ is defined in the following way:

$$AMAE = \frac{1}{J} \sum_{j=1}^{J} MAE_j,$$

  where $AMAE$ values range from 0 to $J - 1$.

- $MMAE$: The Maximum MAE value of all the classes is proposed in this paper as an ordinal regression metric alternative. $MMAE$ is the MAE value of the class with higher distance from the true values to the predicted ones:

$$MMAE = \max\{MAE_j;\ j = 1, ..., J\},$$

where MAE_j is the MAE value for the *j-th* class. $MMAE$ values range from 0 to $J-1$ and it is a natural extension of MS to ordinal regression problems.

Finally association metrics are presented, which are also used in ordinal classification:

- r_S: The Spearman's rank correlation coefficient is a non-parametric measure of statistical dependence between two variables [3]:

$$r_S = \frac{\sum_{i=1}^{n} (O(y_i) - \overline{O(y)})(O(y_i^*) - \overline{O(y^*)})}{\sqrt{\sum_{i=1}^{n} (O(y_i) - \overline{O(y)})^2} \sqrt{\sum_{i=1}^{n} (O(y_i^*) - \overline{O(y^*)})^2}},$$

where $\overline{O(y)}$ and $\overline{O(y^*)}$ are the average of $O(y_i)$ and $O(y_i^*)$, $i = 1, ..., n$, respectively. Recall that $O(C_j) = j$. r_S values range from -1 to 1.

- τ_b: The Kendall's τ is a statistic used to measure the association between two measured quantities. Specifically, it is a measure of rank correlation [4]:

$$\tau_b = \frac{\sum_{i,j=1}^{n} c_{ij}^* c_{ij}}{\sqrt{\left(\sum_{i,j=1}^{n} c_{ij}^{*2}\right)\left(\sum_{i,j=1}^{n} c_{ij}^2\right)}},$$

where c_{ij}^* is +1 if $O(y_i^*) > O(y_j^*)$, 0 if $O(y_i^*) = O(y_j^*)$, and -1 if $O(y_i^*) < O(y_j^*)$ for $i, j = 1, ..., n$, and similar for c_{ij}. The τ_b values range from -1 to 1.

- $WKappa$: The Weighted Kappa is a modified version of the Kappa statistic calculated to allow assigning different weights to different levels of aggregation between two variables [21]:

$$WKappa = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}},$$

where

$$p_{o(w)} = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{J} w_{jk} n_{jk},$$

and

$$p_{e(w)} = \frac{1}{n^2} \sum_{j=1}^{J} \sum_{k=1}^{J} w_{jk} n_{j\bullet} n_{\bullet k},$$

where the weight $w_{jk} = |j-k|$ quantifies the degree of discrepancy between the true (j) and the predicted (k) categories, and $WKappa$ values range from -1 to 1.

While the r_S and τ_b measures are independent on the values chosen for the ranks that represent the classes, MAE, $MMAE$ and $AMAE$ depend on the distance between ranking of two consecutive classes.

2.1. Correlations between metrics

To study the relationships between the different metrics, a correlation matrix comparing them will be analyzed. This matrix is generated from the results obtained by four algorithms from the literature in the ten datasets used in this paper. The information about these datasets can be seen in Section 5. The experimental setup was a stratified 30-holdout in which the training set had approximately 75% of the patterns, and the generalization set had the remaining ones.

The analysis was done for each of the four methods separately in order to take into account the effect of the classifier in the results (Table 2). These matrices were generated in the following way: firstly, since all the methods are deterministic each split of a dataset was run once for each method. So, for each method and dataset, 30 generalization models were available. The correlation between each pair of metrics through the 30 models was calculated leading to a total of 10 correlation matrices (8×8 dimensional matrices). Then, if the correlation value for a given pair of metrics was greater than 0.75, one point was summed to the corresponding pair. Finally, the total points of each pair was divided by the number of comparison (10) to obtain the percentage of times this pair of metrics exhibited a correlation higher than 0.75. When one of the two compared metrics (or both) was constant for the 30 models, the corresponding comparison was ignored, given that correlations could not be obtained. For example, the correlation between CCR and MS with the results of the SVM method is 16.67 due to one comparison was greater than 0.75 and four comparison could not be obtained ($1/6 = 0.1667 = 16.67\%$).

This process was repeated for each method, as can be seen in Table 2. A summary of the four studies has been included in Table 3. This matrix was generated taking into account the comparison of the four methods jointly. As can be seen, the conclusions from the five matrices of both tables are very similar.

The four algorithms are widely used and have been chosen because they usually exhibit good performance. The first one, SVM, is nominal, while others are specific to ordinal regression:

- SVM: Support Vector Machines (SVM) are well known and a robust classification method. In this paper, we use the LibSVM software for the optimization of SVM [22]. This library contains a script for automatically adjusting the hyperparameters associated to this kind of models, including the cost parameter and the width of the Gaussian kernels. The LibSVM grid search cross-

Table 2: Correlation matrices obtained by the different methods. Each element of each matrix is equal to the percentage of times (from a total of 10 comparison) the correlation was higher that 0.75.

SVM method								
	CCR	*MS*	*MAE*	*AMAE*	*MMAE*	r_S	τ_b	*WKappa*
CCR	100.00	**16.67**	80.00	30.00	**10.00**	40.00	60.00	80.00
MS	-	100.00	**16.67**	33.33	33.33	**16.67**	**16.67**	**16.67**
MAE	-	-	100.00	30.00	**10.00**	80.00	80.00	80.00
AMAE	-	-	-	100.00	50.00	50.00	50.00	60.00
MMAE	-	-	-	-	100.00	**20.00**	**10.00**	**10.00**
r_S	-	-	-	-	-	100.00	100.00	90.00
τ_b	-	-	-	-	-	-	100.00	100.00
WKappa	-	-	-	-	-	-	-	100.00
SVMRank method								
	CCR	*MS*	*MAE*	*AMAE*	*MMAE*	r_S	τ_b	*WKappa*
CCR	100.00	*	77.78	33.33	*	33.33	44.44	77.78
MS	-	100.00	*	**16.67**	33.33	*	*	*
MAE	-	-	100.00	22.22	*	88.89	88.89	88.89
AMAE	-	-	-	100.00	55.56	33.33	33.33	44.44
MMAE	-	-	-	-	100.00	**11.11**	**11.11**	**11.11**
r_S	-	-	-	-	-	100.00	100.00	77.78
τ_b	-	-	-	-	-	-	100.00	77.78
WKappa	-	-	-	-	-	-	-	100.00
SVOR-EX method								
	CCR	*MS*	*MAE*	*AMAE*	*MMAE*	r_S	τ_b	*WKappa*
CCR	100.00	40.00	90.00	40.00	**20.00**	40.00	50.00	80.00
MS	-	100.00	40.00	40.00	80.00	40.00	40.00	40.00
MAE	-	-	100.00	40.00	**20.00**	80.00	90.00	80.00
AMAE	-	-	-	100.00	60.00	30.00	50.00	60.00
MMAE	-	-	-	-	100.00	**20.00**	**20.00**	**20.00**
r_S	-	-	-	-	-	100.00	100.00	100.00
τ_b	-	-	-	-	-	-	100.00	100.00
WKappa	-	-	-	-	-	-	-	100.00
SVOR-IM method								
	CCR	*MS*	*MAE*	*AMAE*	*MMAE*	r_S	τ_b	*WKappa*
CCR	100.00	40.00	90.00	40.00	**20.00**	40.00	50.00	80.00
MS	-	100.00	40.00	40.00	80.00	40.00	40.00	40.00
MAE	-	-	100.00	40.00	**20.00**	80.00	90.00	100.00
AMAE	-	-	-	100.00	60.00	50.00	50.00	60.00
MMAE	-	-	-	-	100.00	**20.00**	**20.00**	**20.00**
r_S	-	-	-	-	-	100.00	100.00	90.00
τ_b	-	-	-	-	-	-	100.00	90.00
WKappa	-	-	-	-	-	-	-	100.00

*: means that one of the two compared metrics (or both) was constant for all comparisons and, therefore, the correlation could not be obtained.

Table 3: Matrix summarizing all the correlation matrices of the study. Each element is equal to the percentage of times (from a total of 40 comparison) the correlation was higher that 0.75.

	CCR	*MS*	*MAE*	*AMAE*	*MMAE*	r_S	τ_b	*WKappa*
CCR	100.00	22.73	84.62	35.90	**12.82**	38.46	51.28	79.49
MS	-	100.00	22.73	31.82	54.55	22.73	22.73	22.73
MAE	-	-	100.00	33.33	**12.82**	82.05	87.18	87.18
AMAE	-	-	-	100.00	56.41	41.03	46.15	56.41
MMAE	-	-	-	-	100.00	**17.95**	**15.38**	**15.38**
r_S	-	-	-	-	-	100.00	100.00	89.74
τ_b	-	-	-	-	-	-	100.00	92.31
WKappa	-	-	-	-	-	-	-	100.00

validation procedure has been modified to use *MAE* as the hyper-parameters selection criteria.

- SVMRank: It applies the Extended Binary Classification (EBC) method [23] to SVM. The EBC method can be summarized in the following three steps. First, transform all training samples into extended samples weighting these samples by using the absolute cost matrix. Second, all the extended examples are jointly learned by a binary classifier with confidence outputs, aiming at a low weighted 0/1 loss. Last step is used to convert the binary outputs to a rank.

- SVOR-EX and SVOR-IM: Support Vector Ordinal Regression (SVOR) by Chu and Keerthi [24] proposes two new support vector approaches for ordinal regression. Here, multiple thresholds are optimized in order to define parallel discriminant hyperplanes for the ordinal scales. The first approach includes explicit inequality constraints on the thresholds (SVOR-EX). In the second approach, the samples in all the categories are allowed to contribute errors for each threshold, therefore there is no need of including the inequality constraints in the problem. This approach is named a SVOR with implicit constraints (SVOR-IM).

According to Table 3, the pairs of measures that are less correlated (with a value lower than 20%) are *CCR-MMAE*, *MAE-MMAE*, *MMAE-*r_S, *MMAE-*τ_b and *MMAE-WKappa*. Of these five pairs, the last three are very similar, because the metrics r_S, τ_b and *WKappa* are highly correlated (values higher than 90%). Therefore, in our study will use the pair formed by *MMAE* and τ_b. The choice of τ_b is due to it is one of the most used metrics in ordinal regression besides providing an intuitive view of the results.

Logically, the three selected pairs (*CCR-MMAE*, *MAE-MMAE* and *MMAE-*τ_b) would be ideal objectives to guide the evolution of a multi-objective algorithm, since they have low linear correlations and may implicitly be non-cooperative objectives. In the rest of the paper we will focus on determining which of them presents the most promising results. Note the goal of using MOEA is to optimize objectives which can be non-cooperative in the solution design.

2.2. Comparison of the ordinal metrics

For the metric pairs selected, a study of their behaviour is done by using 6 confusion matrices shown in Table 4. These matrices are designed to cover different extreme possible situations. The last two matrices represent situations where the 50 percent of patterns of

Table 5: Results obtained by the selected metrics.

	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_3$	$\mathbf{M}_4$	$\mathbf{M}_5$	$\mathbf{M}_6$
CCR	0.0	0.5	0.9	0.0	0.5	0.5
MAE	1.0	0.8	0.3	1.0	0.5	1.0
$MMAE$	1.0	2.0	3.0	1.0	0.5	1.0
τ_b	0.1972	0.8280	0.6375	0.4140	0.6669	0.1667

each class are correctly classified. The other patterns are miss-classified in adjacent classes ($\mathbf{M}_5$) or with a classification error of two classes ($\mathbf{M}_6$). All matrices follow the same distribution of patterns per class. The behaviour of the different metrics over other matrices can be seen in [19, 25].

Table 5 shows the results obtained after applying the selected metrics to the confusion matrices. These results show that $\mathbf{M}_1$ and $\mathbf{M}_4$ have similar performance in CCR, MAE and $MMAE$, but not with respect to τ_b. This indicates that τ_b is able to reflect the better performance of $\mathbf{M}_4$ with respect to $\mathbf{M}_1$ (although there the errors are the same, $\mathbf{M}_4$ keeps better the relative order of the patterns, given that patterns of class C_1 are positioned before patterns of class C_2 and patterns of class C_3 before those of class C_4). Analyzing the last two matrices is noted that an increase in the distance of errors produces a degradation of performance of MAE, $MMAE$ and τ_b. In addition, whenever the MAE values of all classes are equal, MAE and $MMAE$ values are identical.

Next, the three selected metric pairs (hereafter referred to as proposals) will be analyzed in order to verify the relationship between the metric of each pair:

- Proposal 1 (CCR-$MMAE$): $\mathbf{M}_3$ has a good value of CCR and a low value of $MMAE$. However, $\mathbf{M}_1$ or $\mathbf{M}_4$ get a $CCR = 0$, but substantially better values of $MMAE$. This indicates that these measures may be non-cooperative.

- Proposal 2 (MAE-$MMAE$): $\mathbf{M}_2$ and $\mathbf{M}_3$ obtain a very acceptable value of MAE, while values of MAE in $\mathbf{M}_1$ and $\mathbf{M}_4$ are worse. This indicates, as in the previous case, that these metrics may be non-cooperative, because in different situations, one of them improves and the other one worsens.

- Proposal 3 (τ_b-$MMAE$): Analysing this proposal, we see that $\mathbf{M}_2$ gets a great value for τ_b, whereas $\mathbf{M}_1$ and $\mathbf{M}_4$ are not getting so good ones. However, $\mathbf{M}_1$ and $\mathbf{M}_4$ obtained better $MMAE$ values than those obtained by $\mathbf{M}_2$. This points out that the measures are non-cooperative.

3. Ordinal model

One main issue of ordinal classification is that there is no notion of the precise distance between classes. The samples are labeled by a set of ranks with different categories and an order. In this paper, the classical Proportional Odds Model (POM) [16] adapted to ANNs [26] is considered. The POM model works based on two elements: the first one is a linear layer with only one node (see Figure 2) whose inputs are stamped onto a line, to give them an order which facilitates ordinal classification. After this one node linear layer, an output layer is included with one bias for each class whose objective is to classify the patterns into their corresponding class. In general, this kind of models are named to as latent variable models or threshold models [27]. The classification rule can be represented in the following general form:

$$C(\mathbf{x}) = \begin{cases} C_1, & \text{if } f(\mathbf{x},\theta) \le \beta_0^1 \\ C_2, & \text{if } \beta_0^1 < f(\mathbf{x},\theta) \le \beta_0^2 \\ \dots & \\ C_J, & \text{if } f(\mathbf{x},\theta) > \beta_0^{J-1} \end{cases}, \tag{1}$$

where the set of biases or thresholds ($\boldsymbol{\beta} = \{\beta_0^1, \dots, \beta_0^{J-1}\}$) satisfy the following ordinal constraint: $\beta_0^1 < \beta_0^2 < \cdots < \beta_0^{J-1}$. $\mathbf{x}$ is the input pattern to be classified, $f(\mathbf{x},\theta)$ is a ranking function and θ is the vector of parameters of the model. Indeed, the analysis of Equation (1) reveals the general idea previously presented: patterns, $\mathbf{x}$, are projected to a real line by using the ranking function, $f(\mathbf{x},\theta)$, and the ordered biases or thresholds, β_0^j, separate the ordered classes.

Let us formally define the model for each class as $f_j(\mathbf{x},\theta,\beta_0^j) = f(\mathbf{x},\theta) - \beta_0^j,\ 1 \le j \le J-1$, where the projection function $f(\mathbf{x},\theta)$ is estimated using S sigmoidal basis functions, $f(\mathbf{x},\theta) = \sum_{s=1}^{S} \alpha_s B_s(\mathbf{x},\mathbf{w}_s)$, replacing $B_s(\mathbf{x},\mathbf{w}_s)$ by the sigmoidal equations:

$$B_s(\mathbf{x},\mathbf{w}_s) = \frac{1}{1+\exp\left(-w_{s0} - \sum_{i=1}^{I} w_{si}x_i\right)}$$

where I is the number of inputs.

By using the POM model [16], this projection can be used to obtain cumulative probabilities:

$$P(Y \le j) = P(Y=1) + \cdots + P(Y=j),\ 1 \le j \le J-1,$$

$$P(Y \le J) = 1,$$

cumulative odds:

$$odds(Y \le j) = \frac{P(Y \le j)}{1 - P(Y \le j)}, 1 \le j \le J-1,$$

Table 4: Confusion matrices evaluated in the study.

$$\mathbf{M}_1 = \begin{pmatrix} 0 & 10 & 0 & 0 \\ 20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 30 \\ 0 & 0 & 40 & 0 \end{pmatrix} \quad \mathbf{M}_2 = \begin{pmatrix} 10 & 0 & 0 & 0 \\ 20 & 0 & 0 & 0 \\ 30 & 0 & 0 & 0 \\ 0 & 0 & 0 & 40 \end{pmatrix} \quad \mathbf{M}_3 = \begin{pmatrix} 0 & 0 & 0 & 10 \\ 0 & 20 & 0 & 0 \\ 0 & 0 & 30 & 0 \\ 0 & 0 & 0 & 40 \end{pmatrix}$$

$$\mathbf{M}_4 = \begin{pmatrix} 0 & 10 & 0 & 0 \\ 0 & 0 & 20 & 0 \\ 0 & 30 & 0 & 0 \\ 0 & 0 & 40 & 0 \end{pmatrix} \quad \mathbf{M}_5 = \begin{pmatrix} 5 & 5 & 0 & 0 \\ 0 & 10 & 10 & 0 \\ 0 & 15 & 15 & 0 \\ 0 & 0 & 20 & 20 \end{pmatrix} \quad \mathbf{M}_6 = \begin{pmatrix} 5 & 0 & 5 & 0 \\ 0 & 10 & 0 & 10 \\ 15 & 0 & 15 & 0 \\ 0 & 20 & 0 & 20 \end{pmatrix}$$

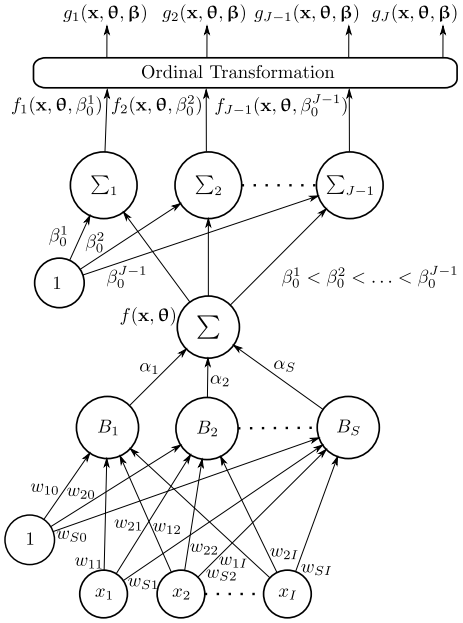

Figure 2: Neural network model for ordinal regression.

and cumulative logits:

$$logit(Y \leq j) = \ln\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) =$$

$$= f(\mathbf{x}, \theta) - \beta_0^j = f_j(\mathbf{x}, \theta, \beta_0^j),$$

$$P(Y \leq j) = \frac{1}{1 + \exp(f(\mathbf{x}, \theta) - \beta_0^j)} = \frac{1}{1 + \exp(f_j(\mathbf{x}, \theta, \beta_0^j))},$$

where $1 \leq j \leq J-1$, $P(Y = j)$ is the probability a pattern $\mathbf{x}$ has of belonging to *j-th* class, $P(Y \leq j)$ is the probability a pattern $\mathbf{x}$ has of belonging to classes 1 to j and the logit is modeled by using the ranking function, $f(\mathbf{x}, \theta)$, and the corresponding bias, β_0^j. We can come back to $P(Y = j)$ from $P(Y \leq j)$:

$$P(y_n^{(j)} = 1|\mathbf{x}, \theta, \beta) = P(Y = j) = g_j(\mathbf{x}, \theta, \beta) =$$

$$= P(Y \leq j) - P(Y \leq j-1), \ j = 1, \ldots, J,$$

where $y_n^{(j)} = 1$ if pattern n belongs to class j and 0 otherwise, and the final probability model can be expressed as:

$$g_j(\mathbf{x}, \theta, \beta) = \frac{1}{1 + \exp(f_j(\mathbf{x}, \theta, \beta_0^j))} - \quad (2)$$

$$- \frac{1}{1 + \exp(f_{j-1}(\mathbf{x}, \theta, \beta_0^{j-1}))}, 1 \leq j \leq J-1,$$

$$g_J(\mathbf{x}, \theta, \beta) = 1 - \frac{1}{1 + \exp(f_{J-1}(\mathbf{x}, \theta, \beta_0^{J-1}))} =$$

$$= \frac{\exp(f_{J-1}(\mathbf{x}, \theta, \beta_0^{J-1}))}{1 + \exp(f_{J-1}(\mathbf{x}, \theta, \beta_0^{J-1}))}.$$

4. Method

To see how the selected metrics behave, this paper uses the MOEA described in [28]. The algorithm used is the Memetic Pareto Differential Evolution Neural Network (MPDENN) algorithm developed by R. Storn and K. Price in [17] and modified by H. Abbass to train neural networks [18]. MPDENN is adapted according to the trade-off between the *CCR* and *MS* analyzed in [12, 29]. The fundamental bases of this algorithm are Differential Evolution (DE) and the concept of Pareto dominance. DE has often been used to train neural networks in the context of both single-objective [30, 31] and multi-objective [13, 32] optimization.

The main feature of the MPDENN algorithm is the inclusion of a crossover operator together with a mutation one. The crossover operator is based on the random choice of three parents, where one of them (main parent) is modified using the weighted difference of the other two (secondary parents). The child generated by the crossover and mutation operators is included in the population if it dominates its main parent, if it has no relationship with him or if it is the best child of the rejected children. A generation of the evolutionary process ends when the population has been completed. At the beginning of each generation, dominated individuals are eliminated from the population. The ordinal metrics are used as fitness functions of the DE algorithm without

Table 6: Characteristics of the datasets.

Dataset	#Patterns	#Attributes	#Classes	Class distribution
Automobile	205	71	6	(3,22,67,54,32,27)
Balance-scale	625	4	3	(288,49,288)
Bondrate	57	37	5	(6,33,12,5,1)
ERA	1000	4	9	(92,142,181,172,158,118,88,31,18)
ESL	488	4	9	(2,12,38,100,116,135,62,19,4)
Eucalyptus	736	91	5	(180,107,130,214,105)
LEV	1000	4	5	(93,280,403,197,27)
SWD	1000	10	4	(32,352,399,217)
Toy	300	2	5	(35,87,79,68,31)
Winequality-red	1599	11	6	(10,53,681,638,199,18)

requiring any change in the evolutionary process. Further details can be found in [28], specially those related to the local optimization procedure included in the DE algorithm. In this paper, the improved Resilient Backpropagation (iRprop$^+$) algorithm [20] is used as the local search procedure. Since the ordinal classification metrics are non-derivable functions, we have selected the cross-entropy error function (E) as a metric to guide the iRprop$^+$ optimization. This metric is proved to be a robust optimization metric for classification problems [33]. The E function is given by:

$$E(\theta, \beta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J} y_n^{(j)} \log g_j(\mathbf{x}_n, \theta, \beta),$$

where $g_j(\mathbf{x}_n, \theta, \beta)$ is the function defined at Equation (2). In the experiments, the local search procedure is applied every five generations.

5. Experimental study

To verify the efficiency of the three proposals, ten ordinal datasets have been used. Nine of them are benchmark datasets[1] and the other (Toy) has been generated following the guidelines in [34]. Table 6 shows the characteristics of the datasets used, including the number of patterns, the number of attributes (after transforming nominal attributes into binary ones), the number of classes and the class distribution (number of patterns for each class).

Due to the fact that the MOEA used is non-deterministic, we perform a stratified 30-holdout, where approximately 75% of the instances are used for the training set and the remaining 25% for the test or generalization set (maintaining the original distribution of classes).

[1]Datasets are available in `http://weka.wikispaces.com/Datasets` and `http://mldata.org/`.

The essential parameters of the algorithm (population size, number of generations and number of nodes in hidden layer) were obtained using a 5-fold cross-validation process on the training set. A grid search was performed using $\{10, 25, 50\}$ for the population size, $\{100, 150, 200\}$ for the number of generations, and $\{5, 10, 20, 30\}$ for the number of nodes in hidden layer. The criterion to select the best parameter combination was the *MAE* metric, due to it is one of the most commonly used ordinal metrics in previous works.

5.1. Comparison methods

The results obtain by a MOEA using the three proposals were compared with the two following related ordinal methods.

- Proportional Odds Model (POM) [16]: this method is a cumulative link model, specifically designed for ordinal regression. This model is inspired by the latent variable motivation which provides a solid probabilistic interpretation. In this work, the POM algorithm is used with the *logit* link function, the most extended one. More details of this method have been seen in Section 3.

- Neural Network based on Proportional Odds Model (NNPOM) [35]: this model is a non-linear version of the POM model, which combines neural networks with a cumulative link model. In this method, the output of a neural network is used as latent variable for the POM model. This type of model can be optimized by maximum likelihood optimization. In this work, the model is optimized using the same local search procedure employed by the MPDENN algorithm, iRprop$^+$. The corresponding parameters have been cross-validated using the *MAE* metric and the same ranges.

5.2. Results

Table 7 shows the *CCR* results obtained after guiding the MOEA with the three proposals. They correspond

to the averages and the standard deviations of the generalization results for the 30 models which are Pareto front extremes generated in 30 runs (one Pareto front for each run is obtained and then the two extremes of the front, in training, are extracted. See Figure 3 to locate these models). In addition, the two ordinal methods are included for comparison. The last part of the Table 7 includes the ranking over all the datasets. The ranking is obtained in the usual manner (for each dataset, a 1 is assigned to the best method, and a 8 to worst one) [36]. Similarly, Tables 8, 9 and 10 show the results obtained for the metrics MAE, $MMAE$ and τ_b, respectively.

From a descriptive point of view, for CCR, the best ranking is obtained by the proposal 2, model 1, and the second by the proposal 1, model 1 (Table 7). For MAE (Table 8), the best rankings obtained are similar to those obtained for CCR. For $MMAE$ (Table 9), the best ranking is obtained by the proposal 3, model 2, and the second by the model 1 of the same proposal. For τ_b (Table 10), the best ranking is obtained by the proposal 1, model 1, and the second by the proposal 2, model 1. These results show that the best proposal is the second one (MAE-$MMAE$ pair), because the trained classifiers has better performance on two metrics (CCR and MAE) and the second best ones on τ_b. It should be pointed out that, according to the experiments, τ_b is not a suitable fitness function since the best and second best mean ranking in τ_b are not achieved by the proposal optimizing τ_b (see Table 10). In addition, the three proposals presented competitive results compared with the reference ordinal methods (POM and NNPOM).

A common feature of the three proposals is that the $MMAE$ models do not perform well for global classification metrics (CCR, MAE and τ_b). The reason for this behavior is that these models are focused on the classification of the worst classified class. However, these models obtain the best results in $MMAE$, minimizing the maximum error across all the classes. In addition, during the evolutionary process, these models help to the opposite extreme models to improve their performance in the worst classified class because they incorporate diversity within the individuals population.

In order to determine the statistical significance of the rank differences observed for each method in the different datasets, a Friedman tests [37] have been carried out with a significance level of $\alpha = 0.05$. When there are significant differences, the Bonferroni-Dunn's test is used to compare all methods to each other. This test considers that the performance of any two methods is deemed to be significantly different if their mean ranks differ by at least the critical difference (CD):

$$CD = q\sqrt{\frac{K(K+1)}{6D}},$$

where K is the number of classifiers, D the number of datasets and the q value can be computed as suggested in [38]. We chose the best performing method (for each metric) as the control one for comparison with the rest. The ranks with significant differences for $\alpha = 0.05$ are marked with an $*$ in Tables 7, 8, 9 and 10 and for $\alpha = 0.10$ with a $\bullet$.

In general, the NNPOM method is the worst one with a significantly worse value for all metrics (with $\alpha = 0.05$ for MAE, $MMAE$ and τ_b and $\alpha = 0.10$ for CCR), when compared with the best method in each case. The POM method is significantly worse in $MMAE$ and τ_b (both with $\alpha = 0.05$) and the proposal 3-, model 2 ($MMAE$ extreme of the τ_b-$MMAE$ pair) is worse in τ_b (with $\alpha = 0.10$). The greatest differences were found in the $MMAE$ results. The reason for these differences may be due to the classical ordinal methods do not consider the performance for the worst classified class.

5.3. Study of the relationship between the MMAE and MAE metrics

In accordance with the results presented in the previous section, a good pair for guiding a multi-objective algorithm could be formed by the MAE and $MMAE$ metrics. The analysis of these metrics is intended to further understand how they guide the DE algorithm. To analyse their relationship we propose the following procedure.

Proposition 5.1. *Let us consider a J-class classification problem. Let MAE and MMAE be respectively the two measures associated with an ordinal classifier g. Without loss of generality, denote the class with maximum MAE_j by $j = J$ (MAE_J). Then:*

$$p_J^* MMAE \leq MAE \leq MMAE, \tag{3}$$

where p_J^* is the estimated prior probability of the J-th class:

$$p_J^* = \frac{n_{J\bullet}}{n}.$$

Proof We begin by proving the upper bound. In general:

$$0 \leq MAE_j \leq MMAE \leq J - 1,$$

so:

$$MAE = \frac{\sum_{j=1}^{J} n_{j\bullet} MAE_j}{n} =$$

$$= \frac{n_{1\bullet} MAE_1 + n_{2\bullet} MAE_2 + ... + n_{J\bullet} MAE_J}{n} \leq$$

Table 7: Results obtained in generalization for the metric CCR: mean and standard deviation ($Mean_{SD}$) and mean ranking ($\overline{R}_{CCR}$).

	Proposal 1 (CCR-$MMAE$)		Proposal 2 (MAE-$MMAE$)		Proposal 3 (τ_b-$MMAE$)			
Dataset	Model 1 (CCR)	Model 2 ($MMAE$)	Model 1 (MAE)	Model 2 ($MMAE$)	Model 1 (τ_b)	Model 2 ($MMAE$)	POM	NNPOM
Automobile	*$60.00_{7.08}$*	*$60.00_{7.08}$*	$59.42_{6.51}$	$59.29_{6.55}$	**$60.13_{5.82}$**	$59.81_{6.03}$	$46.67_{19.42}$	$45.06_{6.24}$
Balance	*$97.24_{1.19}$*	*$97.24_{1.19}$*	**$97.31_{1.06}$**	**$97.31_{1.06}$**	$97.18_{1.55}$	$97.18_{1.55}$	$90.55_{1.86}$	$92.31_{9.96}$
Bondrate	*$51.33_{13.46}$*	*$51.33_{13.46}$*	**$53.11_{8.66}$**	$50.89_{13.16}$	$46.44_{12.68}$	$46.22_{12.74}$	$34.44_{16.05}$	$43.11_{13.42}$
ERA	$27.27_{2.41}$	$25.05_{3.09}$	*$27.55_{2.62}$*	$25.56_{2.94}$	$27.44_{2.55}$	$26.64_{2.59}$	$25.61_{2.11}$	**$27.73_{2.86}$**
ESL	*$70.96_{2.81}$*	$61.17_{17.12}$	$70.05_{3.03}$	$68.14_{8.83}$	**$71.69_{3.34}$**	$68.77_{8.42}$	$70.55_{3.36}$	$65.66_{12.85}$
Eucalyptus	*$59.31_{3.16}$*	*$59.31_{3.16}$*	$57.61_{3.50}$	$57.61_{3.50}$	**$59.64_{3.59}$**	**$59.64_{3.59}$**	$14.93_{1.57}$	$54.04_{4.89}$
LEV	$62.80_{2.55}$	$45.45_{7.87}$	**$62.88_{2.55}$**	$47.05_{6.92}$	*$62.84_{2.31}$*	$45.45_{6.52}$	$62.33_{2.80}$	$62.04_{2.54}$
SWD	$57.19_{3.29}$	$47.65_{7.62}$	**$57.63_{3.04}$**	$48.52_{7.11}$	*$57.32_{3.34}$*	$47.65_{6.74}$	$56.79_{2.96}$	$55.12_{3.42}$
Toy	**$95.78_{2.30}$**	**$95.78_{2.30}$**	*$95.60_{2.60}$*	*$95.60_{2.60}$*	$95.51_{2.61}$	$95.51_{2.61}$	$28.93_{2.55}$	$93.60_{3.33}$
Winequality-red	$59.05_{2.21}$	$42.32_{15.77}$	**$59.94_{1.49}$**	$45.73_{15.81}$	$58.25_{6.43}$	$32.06_{11.01}$	*$59.72_{1.54}$*	$59.49_{2.25}$
$\overline{R}_{CCR}$	*2.95*	5.15	**2.55**	5.15	3.15	5.55	5.80$^\bullet$	5.70$^\bullet$

The best result in CCR is in **bold** face and the second best result in *italics*
$*$ and $\bullet$ stand for significant differences with the Bonferroni-Dunn's test when considering $\alpha = 0.05$ and $\alpha = 0.10$, respectively.

Table 8: Results obtained in generalization for the metric MAE: mean and standard deviation ($Mean_{SD}$) and mean ranking ($\overline{R}_{MAE}$).

	Proposal 1 (CCR-$MMAE$)		Proposal 2 (MAE-$MMAE$)		Proposal 3 (τ_b-$MMAE$)			
Dataset	Model 1 (CCR)	Model 2 ($MMAE$)	Model 1 (MAE)	Model 2 ($MMAE$)	Model 1 (τ_b)	Model 2 ($MMAE$)	POM	NNPOM
Automobile	$0.5212_{0.1153}$	$0.5212_{0.1153}$	$0.5263_{0.1074}$	$0.5269_{0.1073}$	**$0.5186_{0.0914}$**	*$0.5199_{0.0928}$*	$0.9532_{0.6869}$	$0.8513_{0.1501}$
Balance	*$0.0299_{0.0124}$*	*$0.0299_{0.0124}$*	**$0.0297_{0.0129}$**	**$0.0297_{0.0129}$**	$0.0301_{0.0174}$	$0.0301_{0.0174}$	$0.1068_{0.0209}$	$0.1053_{0.1875}$
Bondrate	*$0.6022_{0.1580}$*	*$0.6022_{0.1580}$*	**$0.6011_{0.0961}$**	$0.6467_{0.1812}$	$0.6667_{0.1706}$	$0.6667_{0.1724}$	$0.9467_{0.3206}$	$0.7978_{0.2133}$
ERA	$1.2508_{0.0566}$	$1.3563_{0.1256}$	*$1.2489_{0.0497}$*	$1.3419_{0.1076}$	$1.2657_{0.0622}$	$1.3252_{0.0869}$	**$1.2184_{0.0501}$**	$1.2593_{0.0622}$
ESL	*$0.3049_{0.0299}$*	$0.4557_{0.2669}$	$0.3131_{0.0332}$	$0.3393_{0.1115}$	**$0.2964_{0.0365}$**	$0.3325_{0.1113}$	$0.3104_{0.0380}$	$0.4557_{0.6274}$
Eucalyptus	*$0.4665_{0.0376}$*	*$0.4665_{0.0376}$*	$0.4931_{0.0519}$	$0.4931_{0.0519}$	**$0.4661_{0.0462}$**	**$0.4661_{0.0462}$**	$1.9388_{0.2537}$	$0.5750_{0.0758}$
LEV	*$0.4081_{0.0266}$*	$0.6664_{0.1323}$	**$0.4073_{0.0271}$**	$0.6287_{0.1031}$	$0.4085_{0.0252}$	$0.6533_{0.0973}$	$0.4093_{0.0304}$	$0.4165_{0.0285}$
SWD	$0.4519_{0.0358}$	$0.6067_{0.1266}$	**$0.4455_{0.0327}$**	$0.5933_{0.1071}$	*$0.4489_{0.0367}$*	$0.6049_{0.1016}$	$0.4501_{0.0304}$	$0.4789_{0.0384}$
Toy	**$0.0422_{0.0230}$**	**$0.0422_{0.0230}$**	*$0.0440_{0.0260}$*	*$0.0440_{0.0260}$*	$0.0449_{0.0261}$	$0.0449_{0.0261}$	$0.9809_{0.0389}$	$0.0640_{0.0333}$
Winequality-red	$0.4444_{0.0252}$	$0.8046_{0.3415}$	*$0.4393_{0.0180}$*	$0.7274_{0.3365}$	$0.4622_{0.1072}$	$0.9439_{0.2096}$	**$0.4351_{0.0171}$**	$0.4468_{0.0253}$
$\overline{R}_{MAE}$	*2.85*	5.25	**2.65**	5.15	3.50	5.30	5.20	6.10*

The best result in MAE is in **bold** face and the second best result in *italics*
$*$ and $\bullet$ stand for significant differences with the Bonferroni-Dunn's test when considering $\alpha = 0.05$ and $\alpha = 0.10$, respectively.

Table 9: Results obtained in generalization for the metric $MMAE$: mean and standard deviation ($Mean_{SD}$) and mean ranking ($\overline{R}_{MMAE}$).

	Proposal 1 (CCR-$MMAE$)		Proposal 2 (MAE-$MMAE$)		Proposal 3 (τ_b-$MMAE$)			
Dataset	Model 1 (CCR)	Model 2 ($MMAE$)	Model 1 (MAE)	Model 2 ($MMAE$)	Model 1 (τ_b)	Model 2 ($MMAE$)	POM	NNPOM
Automobile	$1.0319_{0.3119}$	$1.0319_{0.3119}$	$1.0282_{0.2986}$	$1.0282_{0.2986}$	*$0.9914_{0.2363}$*	**$0.9720_{0.2290}$**	$1.8937_{1.6795}$	$2.0135_{0.7245}$
Balance	$0.0813_{0.0608}$	$0.0813_{0.0608}$	**$0.0710_{0.0467}$**	**$0.0710_{0.0467}$**	*$0.0777_{0.0764}$*	*$0.0777_{0.0764}$*	$0.1428_{0.0349}$	$0.2785_{0.3740}$
Bondrate	**$1.9833_{0.6628}$**	**$1.9833_{0.6628}$**	$2.1667_{0.6477}$	$2.1889_{0.6170}$	*$2.0500_{0.4614}$*	*$2.0500_{0.4614}$*	$2.0833_{0.7888}$	$2.5958_{0.7426}$
ERA	$2.1045_{0.2585}$	*$2.0453_{0.2400}$*	$2.1478_{0.2677}$	$2.0719_{0.2946}$	$2.1458_{0.2909}$	**$1.9985_{0.2640}$**	$2.1339_{0.2993}$	$2.2021_{0.5536}$
ESL	$1.2111_{0.4057}$	$1.2795_{0.3970}$	*$1.1156_{0.3596}$*	$1.1668_{0.3866}$	$1.1933_{0.4118}$	$1.1515_{0.4047}$	**$0.9994_{0.3890}$**	$1.5289_{1.1577}$
Eucalyptus	*$0.6886_{0.1065}$*	*$0.6886_{0.1065}$*	$0.7085_{0.1063}$	$0.7085_{0.1063}$	**$0.6699_{0.0831}$**	**$0.6699_{0.0831}$**	$3.6963_{0.6639}$	$0.8502_{0.1489}$
LEV	$1.2294_{0.2431}$	$1.0260_{0.2359}$	$1.2302_{0.2389}$	*$1.0082_{0.2581}$*	$1.2378_{0.2306}$	**$0.9906_{0.2492}$**	$1.3111_{0.2433}$	$1.3095_{0.2070}$
SWD	$0.9117_{0.1108}$	*$0.8620_{0.1397}$*	$0.9500_{0.0843}$	**$0.8581_{0.1663}$**	$0.9375_{0.0715}$	$0.8834_{0.2120}$	$1.1208_{0.1011}$	$1.1083_{0.1729}$
Toy	*$0.1227_{0.0635}$*	*$0.1227_{0.0635}$*	$0.1234_{0.0639}$	$0.1234_{0.0639}$	**$0.1168_{0.0591}$**	**$0.1168_{0.0591}$**	$2.0901_{0.2287}$	$0.1584_{0.0723}$
Winequality-red	$2.0467_{0.2678}$	*$1.9299_{0.4558}$*	$2.1133_{0.3316}$	$2.0583_{0.5272}$	$1.9497_{0.3389}$	**$1.7494_{0.5190}$**	$2.1611_{0.2209}$	$2.2306_{0.2623}$
$\overline{R}_{MMAE}$	4.15	*3.55*	4.80	3.80	3.70	**2.00**	6.40*	7.60*

The best result in $MMAE$ is in **bold** face and the second best result in *italics*
$*$ and $\bullet$ stand for significant differences with the Bonferroni-Dunn's test when considering $\alpha = 0.05$ and $\alpha = 0.10$, respectively.

$$\leq \frac{n_{1\bullet}MMAE + n_{2\bullet}MMAE + ... + n_{J\bullet}MMAE}{n} =$$

$$= MMAE(\frac{\sum_{j=1}^{J} n_{j\bullet}}{n}) = MMAE.$$

On the other hand, the lower bound can be obtained:

$$MAE = \frac{n_{1\bullet}MAE_1 + n_{2\bullet}MAE_2 + ... + n_{J\bullet}MAE_J}{n} \geq$$

$$\geq \frac{n_{J\bullet}MAE_J}{n} = \frac{n_{J\bullet}MMAE}{n} = p_J^* MMAE,$$

since

$$n_{1\bullet}MAE_1 + ... + n_{(J-1)\bullet}MAE_{J-1} \geq 0.$$

Corollary 5.2. *If the dataset is completely balanced,*

Table 10: Results obtained in generalization for the metric τ_b: mean and standard deviation ($Mean_{SD}$) and mean ranking ($\overline{R}_{\tau_b}$).

	Proposal 1 (*CCR-MMAE*)		Proposal 2 (*MAE-MMAE*)		Proposal 3 (τ_b-*MMAE*)			
Dataset	Model 1 (*CCR*)	Model 2 (*MMAE*)	Model 1 (*MAE*)	Model 2 (*MMAE*)	Model 1 (τ_b)	Model 2 (*MMAE*)	POM	NNPOM
Automobile	$\mathbf{0.6730_{0.0869}}$	$\mathbf{0.6730_{0.0869}}$	$\mathit{0.6715_{0.0796}}$	$0.6713_{0.0795}$	$0.6670_{0.0721}$	$0.6661_{0.0727}$	$0.4961_{0.2840}$	$0.3984_{0.1461}$
Balance	$\mathit{0.9718_{0.0114}}$	$\mathit{0.9718_{0.0114}}$	$\mathbf{0.9721_{0.0125}}$	$\mathbf{0.9721_{0.0125}}$	$\mathit{0.9718_{0.0167}}$	$\mathit{0.9718_{0.0167}}$	$0.9015_{0.0194}$	$0.8996_{0.1848}$
Bondrate	$\mathbf{0.4221_{0.2021}}$	$\mathbf{0.4221_{0.2021}}$	$\mathit{0.3620_{0.2084}}$	$0.3491_{0.2415}$	$0.3598_{0.1570}$	$\mathit{0.3620_{0.1616}}$	$0.2897_{0.3017}$	$0.1202_{0.2570}$
ERA	$0.4515_{0.0293}$	$0.4426_{0.0361}$	$\mathit{0.4544_{0.0268}}$	$0.4419_{0.0333}$	$0.4515_{0.0308}$	$0.4455_{0.0350}$	$\mathbf{0.4703_{0.0309}}$	$0.4483_{0.0318}$
ESL	$\mathit{0.8692_{0.0168}}$	$0.8456_{0.0460}$	$0.8681_{0.0159}$	$0.8611_{0.0310}$	$\mathbf{0.8737_{0.0178}}$	$0.8661_{0.0244}$	$0.8661_{0.0189}$	$0.8256_{0.1571}$
Eucalyptus	$\mathbf{0.7514_{0.0227}}$	$\mathbf{0.7514_{0.0227}}$	$0.7346_{0.0341}$	$0.7346_{0.0341}$	$\mathit{0.7510_{0.0287}}$	$\mathit{0.7510_{0.0287}}$	$0.0102_{0.0428}$	$0.6797_{0.0450}$
LEV	$\mathit{0.6480_{0.0255}}$	$0.5675_{0.0635}$	$\mathbf{0.6488_{0.0275}}$	$0.5848_{0.0533}$	$0.6477_{0.0258}$	$0.5768_{0.0569}$	$\mathit{0.6480_{0.0291}}$	$0.6415_{0.0283}$
SWD	$0.5348_{0.0381}$	$0.4442_{0.0911}$	$\mathbf{0.5444_{0.0383}}$	$0.4640_{0.0751}$	$\mathit{0.5404_{0.0409}}$	$0.4495_{0.0728}$	$0.5302_{0.0354}$	$0.5053_{0.0466}$
Toy	$\mathbf{0.9763_{0.0130}}$	$\mathbf{0.9763_{0.0130}}$	$\mathit{0.9749_{0.0155}}$	$\mathit{0.9749_{0.0155}}$	$0.9745_{0.0151}$	$0.9745_{0.0151}$	$-0.0303_{0.1104}$	$0.9635_{0.0181}$
Winequality-red	$0.5034_{0.0309}$	$0.4147_{0.0995}$	$\mathbf{0.5059_{0.0269}}$	$0.4234_{0.1202}$	$\mathit{0.5055_{0.0284}}$	$0.3862_{0.0830}$	$0.5025_{0.0236}$	$0.4885_{0.0319}$
$\overline{R}_{\tau_b}$	**2.40**	4.75	*2.50*	5.25	3.60	$5.60^{\bullet}$	5.30	6.60^{*}

The best result in τ_b is in **bold** face and the second best result in *italics*
$*$ and $\bullet$ stand for significant differences with the Bonferroni-Dunn's test when considering $\alpha = 0.05$ and $\alpha = 0.10$, respectively.

then:

$$n_{1\bullet} = n_{2\bullet} = ... = n_{J\bullet} = \frac{n}{J} \;\; and \;\; p_j^* = \frac{1}{J}.$$

Therefore:

$$\frac{MMAE}{J} \leq MAE \leq MMAE.$$

5.3.1. Graphical representation of MMAE-MAE

The *MMAE-MAE* point of view allows us to represent the performance of a classifier in a two dimensional space, taking into account that this pair of measures consider all the classes of the problem. Concretely, $MMAE$ is represented on the horizontal axis and MAE on the vertical axis. One point in $(MMAE, MAE)$ space dominates another if it is below and to the left, i.e. it has less $MMAE$ and less MAE. Therefore, from the inequality previously derived in (3), each classifier will be represented as a point in the white region in Figure 3. Several points in $(MMAE, MAE)$ space are important to note. The worst classifier is located at the upper right point $(J-1, J-1)$ and the $(0, 0)$ point represents the optimum classifier. In addition, the point $(J-1, p_J^*(J-1))$ corresponds to a classifier that has, at least, one class with the worst classification possible. Note that it is possible to find among them classifiers with a low level of MAE, but with a higher level of $MMAE$, specially when the number of classes is high. Thus, minimizing these two error functions simultaneously produces models which are a trade-off between average results that are acceptable to all classes and the lowest ranked class, i.e. the class that has patterns farthest from the corresponding class in the ordinal ranking.

From the feasible region, the following comments can be made. First of all, observe that a decrease in MAE does not imply a decrease in $MMAE$. Reciprocally, a decrease in $MMAE$ does not mean a decrease in MAE. On the other hand, it should be noted that for a fixed value of MAE, a classifier will be better when it corresponds to a point closer to the diagonal of the $(J-1) \times (J-1)$ square.

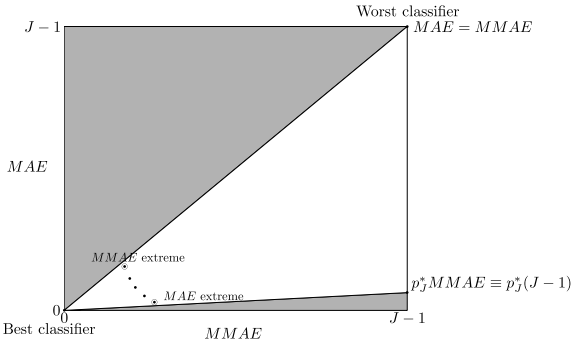

Figure 3: Unfeasible region in the two-dimensional $(MMAE, MAE)$ space.

It is important to analyze if $MMAE$ and MAE are not cooperative in general, especially at certain high levels. At the beginning of a learning process, $MMAE$ and MAE could be cooperative, however after some generations, objectives become non-cooperative and a decrease of one objective usually involves an increase in the other one, as seen in Subsection 2.2.

Figure 4 shows the graphical results obtained with *MAE-MMAE* pair for the SWD dataset. For the *MAE-MMAE* space, the Pareto front for one specific run of the 30 ones performed for each dataset is selected, concretely the execution that presents the best individual on MAE for training data. The test graphic shows the $MMAE$ and MAE values over the generalization set for the individuals who are reflected in the training graphic. Observe that the *MAE-MMAE* values do not form Pareto fronts in generalization, and the individuals

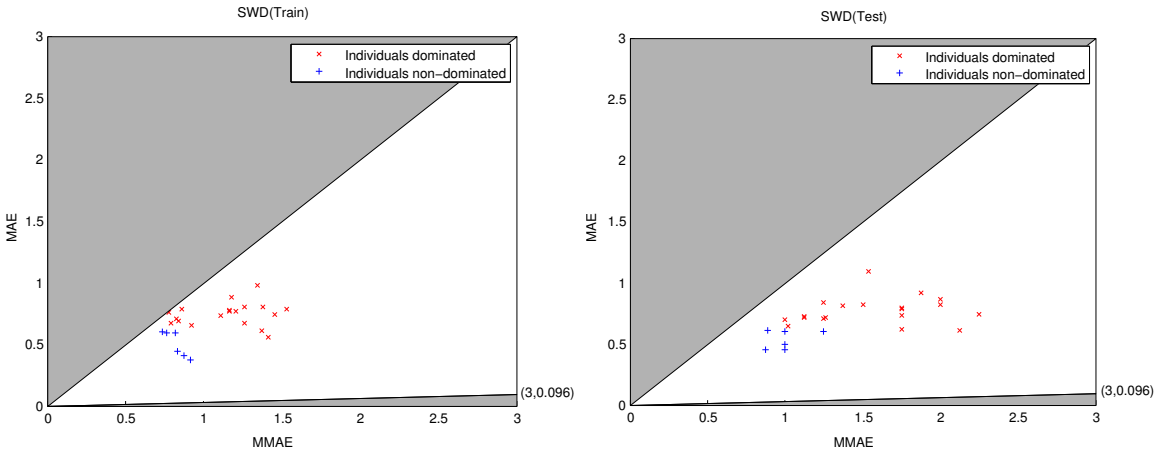

Figure 4: Pareto front in training and associated values in generalization.

that in the training graphic were in the first Pareto front, now can be located within a worst region. In general, the structure of a Pareto front in training has not to be maintained in generalization.

6. Conclusion

This paper contributes an analysis of different state-of-the-art performance measures to evaluate an ordinal classifier. The aim of this analysis is selecting the best pair of metrics to guide a multi-objective evolutionary algorithm. In this analysis, the new $MMAE$ metric is included. This metric is the highest MAE value from MAEs measured independently for each class, i.e. it evaluates the performance of the worst classified class. The analysis studies the correlations between the different metrics for 10 ordinal regression datasets and 4 state-of-the-art methods. Three different pairs of metrics seem to be non-cooperative and, therefore, the most interesting (CCR-$MMAE$, MAE-$MMAE$ and τ_b-$MMAE$). In addition, these measures are studied over a set of synthetic confusion matrices.

To assess these non-cooperative metrics pairs in 10 ordinal datasets, a multi-objective evolutionary algorithm called MPDENN is guided by each of these three combinations. The MPDENN is used to optimize a neural network based on the proportional odds model and the results obtained by the extremes of the Pareto fronts when considering each proposal are reported. These results are compared with those obtained for two reference ordinal methods. This comparison establishes the second proposal (MAE-$MMAE$ pair) as a very competitive one, obtaining suitable classifiers to optimize all the CCR, MAE and τ_b metrics when selecting the MAE extreme of the Pareto front, and acceptable values of $MMAE$ when selecting the $MMAE$ extreme. The reason of this good performance can be found in the fact that a good ordinal classifier must not only classify well the majority classes but also the other classes, including the smallest ones. Finally, the paper analyses the relationship between MAE and $MMAE$ to better understand the 2-dimensional space where the search of the evolutionary algorithm takes place. An inequality is derived, which limits the search space, and some of the Pareto fronts are represented both for training and generalization sets.

Acknowledgment

This work was supported in part by the TIN2011-22794 project of the Spanish Ministerial Commision of Science and Technology (MICYT), FEDER funds and the P08-TIC-3745 project of the "Junta de Andalucía" (Spain). Manuel Cruz-Ramírez's research has been subsidized by the FPU Predoctoral Program (Spanish Ministry of Education and Science), grant reference AP2009-0487. Javier Sánchez-Monedero's research has been funded by the "Junta de Andalucía" Ph. D. Student Program.

References

[1] R. W. Hodge, J. T. Donald, Class identification in the united states, American Journal of Sociology 73 (1968) 535–547.

[2] W. Chu, S. S. Keerthi, New approaches to support vector ordinal regression, in: Proceedings of the 22nd international conference on Machine Learning, 2005, pp. 145–152.

[3] C. Spearman, The proof and measurement of association between two things, American Journal of Psychology 15 (1904) 72–101.

[4] M. G. Kendall, Rank Correlation Methods, New York: Hafner Press, 1962.

[5] L. Goodman, W. Kruskal, Measures of association for cross classifications, Journal of the American Statistical Association 49 (1954) 732–764.

[6] R. H. Somers, The rank analogue of product-moment partial correlation and regression with application to manifold, ordered contingency tables, Biometrika 46 (1955) 241–246.

[7] S. Baccianella, A. Esuli, F. Sebastiani, Evaluation measures for ordinal regression, in: Proccedings of the Ninth International Conference on Intelligent Systems Design and Applications, ISDA'09, 2009, pp. 283–287.

[8] K. Dembczyński, W. Kotlowski, R. Slowiński, Ordinal classification with decision rules, in: Proceedings of the ECML/PKDD'07 workshop on Mining Complex Data, Warsaw, PL, 2007, pp. 169–181.

[9] J. Weston, S. Bengio, N. Usunier, Large scale image annotation: learning to rank with joint word-image embeddings, Mach. Learn. 81 (1) (2010) 21–35. doi:10.1007/s10994-010-5198-3.

[10] J. Weston, S. Bengio, P. Hamel, Multi-tasking with joint semantic spaces for large-scale music annotation and retrieval, Journal of New Music Research 40 (4) (2011) 337–348.

[11] H. Abbass, An evolutionary artificial neural networks approach for breast cancer diagnosis, Artificial Intelligence in Medicine 25 (3) (2002) 265–281.

[12] J. C. Fernández, F. Martínez, C. Hervás, P. A. Gutiérrez, Sensitivity versus accuracy in multi-class problems using memetic Pareto evolutionary neural networks, IEEE Transactions on Neural Networks 21 (5) (2010) 750–770.

[13] J. Sánchez-Monedero, P. A. Gutiérrez, F. Fernandez-Navarro, C. Hervás-Martínez, Weighting efficient Accuracy and Minimum Sensitivity for evolving multi-class classifiers, Neural Processing Letters 34 (2) (2011) 1370–4621.

[14] M. Cruz-Ramírez, C. Hervás-Martínez, J. Fernandez-Caballero, J. Briceño, M. de la Mata, Multi-Objective Evolutionary Algorithm for Donor-Recipient Decision System in Liver Transplants, European Journal of Operational Research 222 (2) (2012) 317–327.

[15] W. Waegeman, B. D. Baets, L. Boullart, ROC analysis in ordinal regression learning, Pattern Recognition Letters 29 (1) (2008) 1–9.

[16] P. McCullagh, Regression models for ordinal data, Journal of the Royal Statistical Society, Series B (Methodological) 42 (2) (1980) 109–142.

[17] R. Storn, K. Price, Differential Evolution. A fast and efficient heuristic for global optimization over continuous spaces, Journal of Global Optimization 11 (1997) 341–359.

[18] H. A. Abbass, R. Sarker, C. Newton, PDE: a Pareto-frontier differential evolution approach for multi-objective optimization problems, in: Proceedings of the 2001 Congress on Evolutionary Computation, Vol. 2, Seoul, South Korea, 2001, pp. 971–978.

[19] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, P. A. Gutierrez, A preliminary study of ordinal metrics to guide a multi-objective evolutionary algorithm, in: 11th International Conference on Intelligent Systems Design and Applications, ISDA 2011, Córdoba, Spain, 2011, pp. 1176–1181.

[20] C. Igel, M. Hüsken, Empirical evaluation of the improved rprop learning algorithms, Neurocomputing 50 (6) (2003) 105–123.

[21] J. L. Fleiss, J. Cohen, B. S. Everitt, Large sample standard errors of kappa and weighted kappa, Psychological Bulletin 72 (5) (1969) 323–327.

[22] C.-C. Chang, C.-J. Lin, LibSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 1–27.

[23] L. Li, H. T. Lin, Ordinal regression by extended binary classification, Advances in Neural Information Processing Systems 19 (2007) 865–872.

[24] W. Chu, S. S. Keerthi, Support Vector Ordinal Regression, Neural Computation 19 (3) (2007) 792–815.

[25] J. S. Cardoso, R. Sousa, Measuring the performance of ordinal classification, International Journal of Pattern Recognition and Artificial Intelligence 25 (8) (2011) 1173–1195.

[26] M. Dorado-Moreno, P. A. Gutiérrez, C. Hervás-Martínez, Ordinal Classification Using Hybrid Artificial Neural Networks with Projection and Kernel Basis Functions, in: 7th International Conference on Hybrid Artificial Intelligence Systems (HAIS2012), 2012, pp. 319–330.

[27] J. Verwaeren, W. Waegeman, B. De Baets, Learning partial ordinal class memberships with kernel-based proportional odds models, Computational Statistics & Data Analysis 56 (4) (2012) 928–942.

[28] M. Cruz-Ramírez, C. Hervás-Martínez, P. A. Gutiérrez, M. Pérez-Ortiz, J. Briceño, M. de la Mata, Memetic Pareto differential evolutionary neural network used to solve an unbalanced liver transplantation problem, Soft Computing 17 (2012) 275–284.

[29] J. C. Fernández, C. Hervás, F. J. Martínez, P. A. Gutiérrez, M. Cruz, Memetic Pareto differential evolution for designing artificial neural networks in multiclassification problems using cross-entropy versus sensitivity, in: Proceedings of the 4th International Conference, HAIS 2009, Vol. 5572, Springer-Verlag Berlin, Heidelberg, 2009, pp. 433–441.

[30] J.-X. Du, D.-S. Huang, X.-F. Wang, X. Gu, Shape recognition based on neural networks trained by differential evolution algorithm, Neurocomputing 70 (4-6) (2007) 896–903.

[31] B. Subudhi, D. Jena, Nonlinear system identification using memetic differential evolution trained neural networks, Neurocomputing 74 (10) (2011) 1696–1709.

[32] M. Cruz-Ramírez, C. Hervás-Martínez, M. Jurado-Expósito, F. López-Granados, A multi-objective neural network based method for cover crop identification from remote sensed data, Expert Systems with Applications 39 (11) (2012) 10038 – 10048.

[33] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[34] J. Pinto da Costa, J. Cardoso, Classification of ordinal data using neural networks, in: Proceedings of the 16th European conference on Machine Learning, ECML'05, Springer-Verlag, 2005, pp. 690–697.

[35] M. Mathieson, Ordinal models for neural networks, in: Neural networks in financial engineering. Proceedings of the 3rd international conference on Neural networks in the capital markets, London, GB, October, 1995, Singapore: World Scientific, 1996, pp. 523–536.

[36] J. Demsar, Statistical comparisons of clasiffiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.

[37] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Annals of Mathematical Statistics 11 (1) (1940) 86–92.

[38] Y. Hochberg, A. Tamhane, Multiple Comparison Procedures, John Wiley & Sons, Inc. New York, NY, USA, 1987.