Open access • Journal Article • DOI:10.1109/TCYB.2013.2266336

Projection-Based Ensemble Learning for Ordinal Regression — Source link

María Pérez-Ortiz, Pedro Antonio Gutiérrez, César Hervás-Martínez

Institutions: University of Córdoba (Spain)

Published on: 01 May 2014 - IEEE Transactions on Systems, Man, and Cybernetics (IEEE Trans Cybern)

Topics: Ordinal regression, Ordinal data, Ensemble learning, Ordinal optimization and Linear discriminant analysis

Projection based ensemble learning for ordinal regression

M. Pérez-Ortiz, P.A. Gutiérrez, *Member, IEEE,* and C. Hervás-Martínez, *Member, IEEE,*

***Abstract*—The classification of patterns into naturally ordered labels is referred to as ordinal regression. This paper proposes an ensemble methodology specifically adapted to this type of problems, which is based on computing different classification tasks through the formulation of different order hypotheses. Every single model is trained in order to distinguish between one given class (k) and all the remaining ones, but grouping them in those classes with a rank lower than k, and those with a rank higher than k. Therefore, it can be considered as a reformulation of the well-known one-versus-all scheme. The base algorithm for the ensemble could be any threshold (or even probabilistic) method, such as the ones selected in this paper: kernel discriminant analysis, support vector machines and logistic regression (all reformulated to deal with ordinal regression problems). The method is seen to be competitive when compared with other state-of-the-art methodologies (both ordinal and nominal), by using six measures and a total of fifteen ordinal datasets. Furthermore, an additional set of experiments is used to study the potential scalability and interpretability of the proposed method when using logistic regression as base methodology for the ensemble.**

***Index Terms*—Ordinal regression, ensemble, discriminant analysis, support vector machines, threshold models, relabelling**

I. Introduction

ORDINAL regression can be defined as a relatively new learning paradigm whose aim is to learn a prediction rule for ordered categories. This problem, firstly arising in statistics [2], is spreading rapidly and receiving a lot of attention from the pattern recognition and machine learning communities [3], [4] because it presents a wide range of applications in areas where human evaluation plays an important role, for example: psychology, medicine, information retrieval, etc. The main difference compared to standard regression is in the target variable, which is composed of finite and discrete category labels, the distances between them being unknown. Concerning classification, the variable to predict is not numerical or nominal, but ordinal; thus these categories show an implicit and natural order. An explanatory example of order among categories could be the Likert scale, a well-known methodology used for questionnaires, where the categories correspond to the level of agreement or disagreement with a series of given statements. The scheme of a typical five-granularity Likert scale could be: {*Strongly disagree, Disagree, Neither agree or disagree, Agree, Strongly Agree*}, where the natural order among categories can be appreciated. The major problem within this kind of classification is that misclassification errors should not be treated equally: misclassifying the *Strongly disagree* class as *Strongly agree* should be more penalized than misclassifying it as *Disagree*. Therefore, several issues must be taken into account in order to exploit the presence of this order among categories. Firstly, this implicit data structure should be learnt by the classifier in order to minimize ordinal classification errors and, secondly, several measures or metrics should be developed in order to do so, given that simply being accurate might not be enough for this kind of problems.

Several approaches to tackle ordinal regression have been proposed in the domain of machine learning over the years, since the first work dating back to 1980 [2]. The simplest idea is to transform these ordinal scales into numeric values and solve the problem as a standard regression one. Kramer et al. investigated and proposed the use of a regression tree learner in this sense [5]. However, as outlined before, there is an important problem within these approaches: the fact that, in general, there is no knowledge about the distances between different classes. On the other hand, other works focused on addressing the problem by simply performing multinomial classification tasks (totally forgetting the order information) or by considering cost-sensitive classification [6] based on trivially imposed cost matrices. Some researchers approach the problem by decomposing the original ordinal regression task into a set of binary classification tasks [3], [7], or by formulating the original problem as one of extended binary classification [8], [9]. However, the most popular approach is clearly the use of threshold models [4], [10]–[12]. These methods are based on the idea that, in order to model ordinal classification problems from a regression perspective, one can assume that some underlying real-valued outcomes exist (also known as latent variable), although they are unobservable. Consequently, these methodologies estimate:

- A function $f(\mathbf{x})$ that tries to predict the nature of those underlying real-valued outcomes.
- A set of bias terms or thresholds $\mathbf{b} = (b_1, b_2, \ldots, b_{K-1}) \in \mathbb{R}^{K-1}$ (where K is the number of classes in the problem) to represent the intervals in the range of $f(\mathbf{x})$, where $b_1 \leq b_2 \leq \ldots \leq b_{K-1}$.

Nowadays, the ensemble paradigm is one of the most actively researched in pattern recognition and machine learning [13]. This methodology imitates human nature to seek several opinions before making a crucial decision [14] and was

M. Pérez-Ortiz, P.A. Gutiérrez and C. Hervás-Martínez are with the Department of Computer Science and Numerical Analysis, University of Córdoba, Campus de Rabanales, Albert Einstein building, 14071 - Córdoba, Spain, e-mail: {i82perom, pagutierrez, chervas}@uco.es.

This paper is a very significant extension of [1] with much additional material, including a comprehensive review of some ordinal regression methodologies, a more detailed description of the proposal with some changes, and a wider experimental section, where the results for different benchmark datasets and measures were analyzed. Besides, SVMs and logistic regression techniques formulated for ordinal regression were also considered in this work, both for ensemble construction and for comparison.

proposed as an alternative to the conventional "standalone" methods, which can be suboptimal. The main aspects addressed in ensemble literature are: development of methods for reducing the dependence between classifiers, i.e. maximizing diversity, and development of effective combination rules.

This paper contributes a novel and natural ensemble methodology to tackle ordinal information which could be used with any threshold model as base classifier. More specifically, in this paper kernel discriminant analysis (KDA) [4], [15] and support vector machines (SVM) [16], [17] were used for a first set of experiments, since these can be considered accurate and successful methods when adapted to ordinal regression [18], [19]. Moreover, logistic regression (LR) [2], [20] was considered for a set of large-scale datasets. The main motivation is the development of an ordinal ensemble algorithm which could benefit from the order information of the data to improve the performance of other existing techniques. As many classifiers as the number of classes are trained, and each single model is computed to differentiate each class from the remaining ones taking ordinal ranks into account, i.e. separating each class from the previous and following classes. The ensemble methodology proposed is based on decomposing ordinal regression problems into simpler classification tasks, where the order information is explicitly included. For a K class ordinal regression problems, 2 binary classification problems and $K-2$ ordinal ones (each composed of three classes) are derived, in such a way that the main classification problem is simplified. This procedure can be appreciated in Fig. 1 for a 5 classes example. The main hypothesis is that the performance of any ordinal algorithm could be improved by simplifying classification tasks and formulating multiple order hypotheses which will be combined in a final decision function. The proposal can be seen as a reformulation of the one-versus-all idea to tackle ordinal regression. A set of experiments is presented in this paper, which tests and validates this methodology and other nominal and ordinal ones, taking into account 15 datasets with different characteristics. The results suggest that the proposal reaches a competitive performance level and is able to extract better quality classifiers from the order information in the class labels. Finally, a different set of experiments over two large-scale datasets is conducted to analyze the potential scalability and interpretability of the proposed ensemble.

Some advantages and decisions related to the proposal are now discussed. First of all, the choice of threshold models as base classifiers is justified because of their inherent advantage to lend themselves to probabilistic outputs, as these conditional probabilities of class membership are useful for constructing a more robust ensemble methodology. The proposal can be applied to any threshold model (indeed to any algorithm leading to probabilistic outputs), since the main idea is to compute one model to differentiate each class from the rest by taking ordinal ranks into account, and then extracting final output probabilities from the outcomes of each model. In addition, threshold methods depend to a great extent on the bias or threshold computation, which may be a complex handicap when dealing with kernel methods because of their tendency to over-fit. Instead of using crisp values, this study considers probability estimations to relax and alleviate the misclassification error of multiple order hypotheses. On the other hand, selecting the number of classifiers has always been one of the most important and controversial issues in the ensemble paradigm (this value is usually assigned to an odd number in order to avoid draws), but in this case it is very intuitive, as the number of classifiers would be preassigned to the number of classes in the sample. Also, inducing diversity in the classifiers is a crucial ingredient for developing robust ensemble techniques. However, in this case diversity is implicit in the technique, as each computed model will be composed of different data labelling and pattern distributions. Finally, the proposal could also be justified by the low number of ordinal ensemble methods existing in the literature.

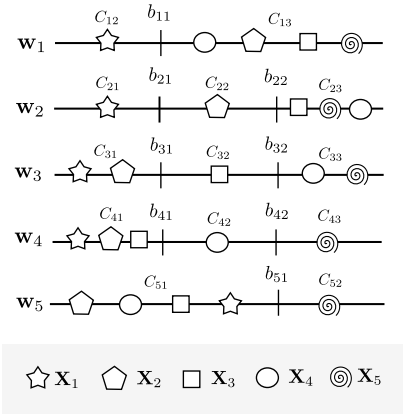

Fig. 1. Example showing different projections computed for the ensemble when $K = 5$. $\mathbf{X}_i$ are the patterns associated to class i. The model trained for separating class i-th from the remaining ones is denoted by $\mathbf{w}_i$ and the corresponding thresholds associated by b_{i1} and b_{i2}. $\mathcal{C}_{ij}$ is used for denoting a synthetically constructed cluster of classes for decision maker i-th.

The paper is organized as follows: Section II shows a description of the methodologies used for the ensemble; Section III formally presents the proposal of this work; Section IV describes the characteristics of the datasets and the experimental study; Section V analyzes the results obtained; and finally, Section VI outlines some conclusions and future work.

II. Previous Notions

In this section, the terminology and notation that will be used throughout the entire work is established. The goal in classification is to assign an input vector $\mathbf{x}$ to one of K discrete classes $\mathcal{C}_k$, where $k \in \{1, \ldots, K\}$. Thus, a formal framework for the ordinal regression problem could be introduced by considering an input space $X \in \mathbb{R}^d$, where d is the data dimensionality. To do so, an outcome space $Y = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ is defined, where the labels are ordered due to the data ranking structure ($\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_K$, where $\prec$ denotes this order information). Let N be the number of patterns in the sample and N_k the number of samples for the k-th class. The objective in this kind of problem is to find a prediction function $f : X \to Y$ by using an i.i.d. sample $D = \{\mathbf{x}_i, y_i\}_{i=1}^N \in X \times Y$.

The ensemble approach here proposed is applied to three well-known techniques: KDA, SVM and LR. Since they have

been reformulated to deal with ordinal regression problems a brief explanation of these methods is included in this section.

A. Kernel discriminant learning

This learning paradigm (KDA) is one of the pioneer and leading techniques in the machine learning area, since it dates back to 1936 and has been widely used as much for supervised dimensionality reduction as for classification [21]. KDA has also been adapted to ordinal classification [4] by imposing a constraint on the projection to be computed, so that it will preserve and take advantage of the ordinal information from different classes. The method is known as kernel discriminant learning for ordinal regression (KDO) [4].

B. Support vector machines

The SVM paradigm [16], [22] is considered the most common kernel learning method for statistical pattern recognition. This study considers two of the most commonly used approaches for solving multiclass problems with SVMs: the one-vs-all formulation and the one-vs-one formulation.

Some works in the SVM literature have been focused on the reformulation of this successful paradigm to tackle ordinal regression problems [17], [23], [24]. All these approaches share one common objective which is the definition of $K-1$ discriminant hyperplanes represented by the vector $\mathbf{w}$ and the scalars bias $b_1 \leq \ldots \leq b_{K-1}$ in order to properly separate training data into ordered classes by modeling ranks as intervals on the real line.

The proposal of Herbrich [23] derived the well-known SVM methodology for ordinal regression by making use of an independent distribution model and inducing an ordering in the space $\mathbf{X}$ that incurs the smallest number of inversions on pairs $(\mathbf{x}_i, \mathbf{x}_j)$ of objects, the probability of that incurred inversion being given by a risk function for each pair of ranks. The main disadvantage of this algorithm is that the problem is formulated as a quadratic function directly depending on the training number of patterns.

On the other hand, the work of Shashua and Levin [24] introduced two different approaches: the former tries to maximize the margin between the closest neighboring classes by applying the "fixed margin" policy and the latter allows for different margins where the sum of margins is maximized. The principal disadvantage of their proposal is that ordinal inequalities on the thresholds, $b_1 \leq b_2 \leq \ldots \leq b_{K-1}$, are not included in the formulation and this omission may result in disordered thresholds at the solution.

A third proposal of SVMs for ordinal regression is presented in the work of Chu and Keerthi [17]. This study also shows two different implementations for the idea. Both approaches guarantee that the thresholds are properly ordered at the optimal solution. The first one only takes into account adjacent ranks for the determination of the thresholds, whereas in the second one, the whole training sample considering all ranks is used for the determination of each threshold, and samples in all the categories are allowed to contribute errors for each hyperplane. This second approach is called support vector ordinal regression with implicit constraints (SVOI).

From another point of view, ordinal regression can be transformed into several binary classification problems; one binary classifier can be derived for each problem, and the output of all classifiers can be combined to obtain a final decision. The strategy is based on simply checking if the rank of a pattern is greater than a given rank k, $1 \leq k \leq K-1$, which is indeed a binary classification question which is answered by each classifier. This approach is closely related to that proposed in this paper and was first presented in the work of Frank & Hall [3] with C4.5 classification trees as base classifiers. However, SVMs have performed very competitively for binary problems, and a similar proposal was then considered for SVMs in the work of Waegeman & Boullart [7], but introducing specific weights into the different patterns. These weights try to reflect the fact that not all patterns in the "greater than k" class (for the binary classifier k) are equally far from k in the ordinal scale, and they should be treated differently when constructing the classifier (even though they belong to the same class). Both methods will be considered in the experimental section.

C. Logistic regression

In machine learning, LR [20] is a well-known methodology based on a regression analysis for classification problems. This method has been reformulated to deal with ordinal problems giving rise to the proportional odds model (POM) [2]. This model was the first threshold method applied to ordinal regression problems and it is based on a linear projection jointly trained with a set of thresholds by using a similar technique to that considered for nominal LR. Let h denote an arbitrary monotonic link function. The model:

$$h\left(P(y \leq \mathcal{C}_j|\mathbf{x})\right) = \mathbf{w}^\top \mathbf{x} - b_j, \quad j = 1, \ldots, K-1, \qquad (1)$$

links the cumulative probabilities to a linear predictor and imposes an stochastic ordering of the space $\mathcal{X}$, where b_j is the threshold separating $\mathcal{C}_j$ and $\mathcal{C}_{j+1}$ and $\mathbf{w}$ is a linear projection.

III. Ensemble learning for ordinal regression (ELOR)

In the previous section, three well-known classification methods have been presented: KDA, SVM and LR. These methods share one common and general objective which defines the optimization function: the maximization of the distance between different classes. Therefore, they depend greatly on the number of classes in the sample, hindering the separation between them when this number is high. Because of that, the proposed methodology tries to simplify the task of classification, and thus the optimization process. The proposal is intended to construct an ensemble which performs much simpler classification tasks. In order to do so, different decision models are computed, one for separating each class from the remaining ones (avoiding the problem of a great number of classes and aiming at a more balanced classification). The main motivation for this work could be found in the sentence of Albert Einstein, *"Make everything as simple as possible, but not simpler"*, because the original classification

task is simplified, but without forgetting the ordinal ranking information implicit in the data.

Various supervised and disjoint clusters (the term cluster is used to refer to a group of classes) are computed and classified taking into account the natural order of the classes, i.e. a label manipulation procedure is conducted in order to generate multiple hypotheses. In methods that manipulate the target attribute, instead of inducing a single complex classifier, several classifiers are induced with different and usually simpler representations of the target attribute [14]. One example of this is the one-versus-all methodology [25] (previously introduced for SVMs), where a K class classification problem is transformed into K binary classification ones. The one-vs-all paradigm seeks the i-th decision function $f_i(\mathbf{x}), i \in \{1,\ldots,K\}$ fulfilling that $f_i(\mathbf{x}) > 0$ when $\mathbf{x}$ belongs to class i, and $f_i(\mathbf{x}) < 0$ when $\mathbf{x}$ belongs to one of the remaining classes. Therefore, f is used as a membership function for choosing the final prediction. The proposal described in this section can be seen as a one-versus-all reformulation for ordinal regression.

In ordinal regression, one-vs-all approach would not compute a fair classification, as the implicit order information would be ignored. For example, for a 5-class problem, f_4 will try to distinguish between class 4 and classes $\{1,2,3,5\}$. As class 5 is supposed to be closer to class 4 than to classes $\{1,2,3\}$, it might be difficult to separate it from class 4. The proposal tries to separate one class from the previous and the following ones, in such a way that the order among the classes is taken into account (see Fig. 1).

Furthermore, there exists another main issue apart from the exploitation of ordinal ranks by simplifying the classification task. It is well-known that the possible ways of combining the outputs of different classifiers in an ensemble depends on what information is obtained from individual members. When dealing with classification algorithms, the most common output for a learning procedure is the label predicted. However, in some cases, there is other information directly extractable from the classifier which may be helpful for improving classification performance, such as predicted probabilities. Threshold methods present the problem of threshold computation which may often be a complex but important issue, as final classification entirely depends on those thresholds. In order to relax and alleviate this kind of errors, probability estimations are carried out by the proposed ensemble methodology.

Let us formally define the method. Given K different classes and corresponding events $(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K)$, K different classification problems will be computed by relabelling the data and training the learning algorithm with these relabelled patterns. By doing this, K different models will be obtained:

- Two of the models (the first one, $i = 1$, and the last one, $i = K$) will compute binary classifications, separating class i from all the others. Standard KDA, SVM or LR will be applied in these cases.
- The rest of them ($i \in \{2,\ldots,K-1\}$) will be three class classifiers, separating the corresponding class i-th from previous ones $(1,\ldots,i-1)$ and subsequent ones $(i+1,\ldots,K)$. Any of the previously presented ordinal algorithms could be used in order to maintain the ordinal rank of the classes (in these cases, the KDO, SVOI and POM algorithms will be used).

An ensemble set $\mathbb{D}$ will be defined consisting of a combination of K different decision makers, $\mathbb{D} = \{D_1,\ldots,D_K\}$. Each projection will be determined by the set of data to discriminate, as can be seen in Fig. 1 for $K = 5$, where X_i is the set of patterns belonging to class i-th.

The training set is defined as $\mathbb{G} = \{\mathbf{G}_1,\ldots,\mathbf{G}_K\}$ for each member of the ensemble, where $\mathbf{G}_i = \{X_{(j|j<i)}, X_{(j|j=i)}, X_{(j|j>i)}\}$. Note that, in the first and last cases, one of the sets to discriminate will be the empty set, as there are no lower and higher ranking classes, respectively. Consequently, the cardinality of $\mathbf{G}_i$ will be $|\mathbf{G}_i| = 3$, for $i \in \{2,\ldots,K-1\}$, and $|\mathbf{G}_i| = 2$ for $i = 1$ and $i = K$.

Clusters grouping different classes will be defined for each decision maker D_i: $\mathcal{C}_{ij}$, $1 \leq i \leq K$. The set of events to classify is defined in the following way: $\{\mathcal{C}_{i1} = (\mathcal{C}_1 \cup \ldots \cup \mathcal{C}_{i-1}), \mathcal{C}_{i2} = \mathcal{C}_i, \mathcal{C}_{i3} = (\mathcal{C}_{i+1} \cup \ldots \cup \mathcal{C}_K)\}$, taking into account that, in the first and the last classification tasks, some of them will be the empty set. These clusters result in different class targets (according to their rank): $S_1 = \{1,2\}$, $S_i = \{1,2,3\}$, $(1 < i < K)$, and $S_K = \{1,2\}$.

Then, each decision maker (D_i) is determined by the set to discriminate $(\mathbf{G}_i)$, the labels S_i, the computed optimal model (which in this case will be the optimal projection or hyperplane $\mathbf{w}_i$) and the set of thresholds for separating the classes $(\mathbf{b}_i)$. Note that the number of thresholds for the classification corresponds to $|S_i| - 1$.

Although KDA, SVM and LR have been selected as base methods since they can be easily transformed to predict probabilities, the ensemble could be used with any threshold or probabilistic method. As when using threshold models it is possible to estimate K sets of probability, the first hypothesis is that the true values of $P(\mathcal{C}_i|\mathbf{x},\mathbb{D})$, i.e. the posterior probability, are the ones most agreed upon by the ensemble.

Although many types of uncertainty exist, probabilistic models fits surprisingly well in most pattern recognition problems [13]. Because of that, this paper tries to construct a classifier by only taking estimated probabilistic information into account. For each pattern and decision maker i, the probability of belonging to class i will be calculated, along with the probability of belonging to the previous classes and the probability of belonging to the following ones. Then, a methodology for joining all the probabilities is proposed. For that, there are several issues to be addressed:

1) Distributing the probabilities within the cluster: when the specific model for separating class i-th from the rest is computed, three (or two) different supervised clusters are formed, one for the classes whose class target is less than i, one for class i and one for the classes whose class target is greater than i. These projections can be used to approximate the probability of belonging to a specific cluster (by using equations (2) and (3) of the next subsection), where one or more classes are represented. This probability has to be distributed among the different classes included in the cluster to obtain a K-class probability distribution for each decision maker.
2) Combining the probabilities: as in any ensemble, a way has to be selected to combine the decisions of all

classifiers (average, product, majority voting, etc).

3) Weighting more prominent classes: after distributing the probabilities, there are classes that are more prominent (for example extreme classes, which appear isolated in two of the projections, see Fig. 1). If a weighting method is not applied, all the patterns will be more likely to be classified in these classes.

A. Obtaining probability outputs

An important advantage of threshold methods [4], [10] over other algorithms is that their outputs can be easily transformed into conditional probabilities by analysing projected patterns and the corresponding thresholds. This is due to the fact that, in high-dimensional feature space, the histogram of each class projected by the discriminant function can be closely approximated by a given distribution. For example, given a pattern $\mathbf{x}$ and a decision maker D_i the probability that this pattern has of belonging to cluster $\mathcal{C}_{ij}$ can be estimated using:

- The probit function, which computes a normal cumulative distribution:

$$P(\mathcal{C}_{ij}|\mathbf{x}, D_i) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mathbf{x}} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt, \qquad (2)$$

- or the logit function, which computes the standard logistic sigmoid:

$$P(\mathcal{C}_{ij}|\mathbf{x}, D_i) = \frac{1}{1+e^{-t}}, \qquad (3)$$

where $i \in \{1, \ldots, K\}$, $j = 1$ or $j \in \{1, 2\}$, $t = \mathbf{w}_i^T\mathbf{x} - b_{ij}$ is the projected pattern, $\mathbf{w}_i^T$ is the i-th transposed projection vector, b_{ij} is the corresponding bias for cluster j, and the assumption of $\mu = 0$ and $\sigma = 1$ is made. Conditional probabilities can be useful, for instance, in applications where the output of a classifier needs to be combined with other information, and it is not only the class assignment that is interesting, but also its probability. Additionally, these probabilities allow us to combine the outputs of K classifiers.

In this work, the probit function has been used for estimating the probabilities in the case of the KDA methodologies, since these methods assume an unimodal normal distribution on the data. For LR methods, the logit function was used. On the other hand, as there is no guideline about which method should be used with nonparametric methods, such as SVMs, the logit function has been considered, which has been proved to show good results with this technique [26], [27].

B. Distributing the probabilities within the cluster

If the probability that a pattern belongs to a specific cluster is determined by a decision maker D_i, then when the cluster $\mathcal{C}_{ij}$ has only one class, the probability is directly defined but, if there are multiple classes, this probability should be distributed among the classes included in it (as can be seen in Fig. 2). One first idea could be simply to ignore all the clusters with more than one class and make use of the independent membership values of the i-th single class of each decision maker (after applying the transformations proposed in the previous subsection), in such a way that a vector of decision values $\mathbf{V} = \{P(\mathcal{C}_1|\mathbf{x}, D_1), P(\mathcal{C}_2|\mathbf{x}, D_2), \ldots, P(\mathcal{C}_K|\mathbf{x}, D_K)\} = \{P(\mathcal{C}_{11}|\mathbf{x}, D_1), P(\mathcal{C}_{22}|\mathbf{x}, D_2), \ldots, P(\mathcal{C}_{K2}|\mathbf{x}, D_K)\}$ is computed and the final prediction would be the index of the maximum value of it. Throughout this work, this methodology is referred to as simple ensemble learning for ordinal regression (SELOR) and has a significant disadvantage: the whole set of probabilities is not being considered.

More complex responses can be obtained if clusters with multiple classes are considered and the corresponding probability is distributed among these classes. One possible way of distributing these probabilities is the following:

$$P(\mathcal{C}_k|\mathbf{x}, D_i) = P(\mathcal{C}_{ij}|\mathbf{x}, D_i) \cdot \gamma_{ik}, \quad \forall(\mathcal{C}_k \in \mathcal{C}_{ij}), \qquad (4)$$

with $k \in \{1, \ldots, K\}$, $j \in \{1, 2, 3\}$ or $j = \{1, 2\}$, and taking into account that $\gamma_{ik} = 1$ when $|\mathcal{C}_{ij}| = 1$.

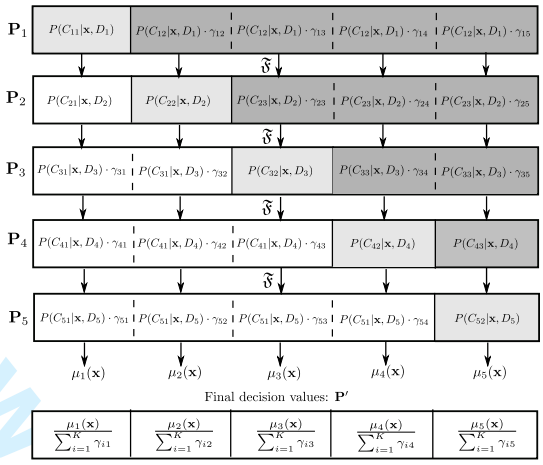

Fig. 2. Example showing the different stages of the procedure. A combination function $\mathfrak{F}$ is used to combine the probability outputs and obtain all $\mu_k(\mathbf{x})$ values.

This γ_{ik} weighting parameter could be chosen in many different ways:

1) Equally distributed probabilities: The probability of belonging to class $\mathcal{C}_k$ for a specific decision maker D_i (where $k \in \{1, \ldots, K\}$) is the probability of belonging to the cluster $\mathcal{C}_{ij}$ (taking into account that the patterns $\mathbf{X}_k$ associated to $\mathcal{C}_k$ belongs to cluster $\mathcal{C}_{ij}$) divided by the number of different class targets involved in the cluster, in this case:

$$\gamma_{ik} = \frac{1}{|\mathcal{C}_{ij}|}, \quad \forall(\mathcal{C}_k \in \mathcal{C}_{ij}).$$

For the sake of simplicity, this will be the method considered for all the experiments in this paper.

2) Distribution according to the number of patterns in each class: The probability of belonging to class $\mathcal{C}_k$ for a decision maker D_i would be the probability of belonging to cluster $\mathcal{C}_{ij}$ multiplied by the number of patterns in class $\mathcal{C}_k$ with respect to the total involved in the cluster,

then:

$$\gamma_{ik} = \frac{\sum_{n=1}^{N} I(y_n = \mathcal{C}_k)}{\sum_{n=1}^{N} I(y_n \in \mathcal{C}_{ij})}, \quad \forall(\mathcal{C}_k \in \mathcal{C}_{ij}),$$

where $I(\cdot)$ is defined as the indicator function.

3) Distribution according to the inverse of the number of patterns of each class: The probability of belonging to class $\mathcal{C}_k$ for a decision maker D_i would be, as before, the probability of belonging to cluster $\mathcal{C}_{ij}$ multiplied by the inverse of the number of patterns in class $\mathcal{C}_k$ with respect to the total involved in the cluster, thus:

$$\gamma_{ki} = 1 - \frac{\sum_{n=1}^{N} I(y_n = \mathcal{C}_k)}{\sum_{n=1}^{N} I(y_n \in \mathcal{C}_{ij})}, \quad \forall(\mathcal{C}_k \in \mathcal{C}_{ij}).$$

This alternative method could be considered for those unbalanced datasets where there is a special interest in classifying minority classes.

Note that the parameter γ_{ki} is calculated taking into account only training data.

C. Fusion of probabilities

After applying the method in the above subsection, a matrix $\mathbf{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_K\}$ of probabilities is obtained, where $\mathbf{P}_{i,j} = p_{i,j} = P(\mathcal{C}_j|\mathbf{x}, D_i)$, satisfying that $\sum_{j=1}^{K} p_{i,j} = 1$. Now, all the columns of this matrix are combined to obtain a final decision vector. A "nontrainable" combiner [13] is considered, i.e. no additional parameters will be tuned, so the ensemble will be ready for classification as soon as the base classifiers are trained. The membership for the j-th class is calculated using the j-th column of the matrix: $\mu_j(\mathbf{x}) = \mathfrak{F}[p_{1,j}(\mathbf{x}), \ldots, p_{K,j}(\mathbf{x})]$, where $\mathfrak{F}$ is defined as a combination function. The most commonly used choices for this function are the simple mean:

$$\mu_j(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^{K} p_{i,j}(\mathbf{x}), \tag{5}$$

and the product:

$$\mu_j(\mathbf{x}) = \prod_{i=1}^{K} p_{i,j}(\mathbf{x}). \tag{6}$$

A theoretical framework is offered for the average and product combiners in [28] based on the Kullback-Leibler divergence, which measures the distance between two probability distributions. These combiners are the two most studied [29], but there is no guideline as to which one is better for a specific problem. In general, the average might be less accurate than the product for some problems, but it is more stable since a small change in a probability makes a bigger impact on the product than on the average.

D. Weighting more prominent classes

The distribution of probabilities considered in subsection III-B makes some classes receive more attention: for example, in Fig. 1, classes $\mathcal{C}_1$ and $\mathcal{C}_5$ appear isolated in the projections more often than classes $\mathcal{C}_2$, $\mathcal{C}_3$ and $\mathcal{C}_4$, their computed probability being higher (a priori) than that of the other classes. Therefore, a weighting method is used in such a way that:

$$P'(\mathcal{C}_i|\mathbf{x}, \mathbb{D}) = \frac{P(\mathcal{C}_i|\mathbf{x}, \mathbb{D})}{\sum_{j=1}^{K} \gamma_{ij}}. \tag{7}$$

E. Further considerations

In order to clarify all the concepts in previous subsections, a summary of the approach in this work is given in Fig. 3.

Pseudocode for the ordinal ensemble proposed

- **Input**: training inputs ($\mathbf{x}_{\mathrm{Tr}}$), training targets ($\mathbf{t}_{\mathrm{Tr}}$), test inputs ($\mathbf{x}_{\mathrm{Ts}}$).
- **Output**: test predicted targets ($\mathbf{t}_{\mathrm{Ts}}$).

for $i = 1$ to K

1. Compute the clusters $\mathbf{G}_i$ from $\mathbf{x}_{\mathrm{Tr}}$ and $\mathbf{t}_{\mathrm{Tr}}$, where $\mathbf{G}_i = \{X_{(j|j<i)}, X_{(j|j=i)}, X_{(j|j>i)}\}$.
2. Train decision maker D_i for $\mathbf{G}_i$: optimal projection $\mathbf{w}_i$ and thresholds ($\mathbf{b}_i$) using either the binary or ordinal algorithm.
3. Project test data.
4. Compute test probabilities of belonging to each cluster.
5. Distribute clustered test probabilities among the classes for obtaining $\mathbf{P}$, equation (4).

end for

Apply the defined $\mathfrak{F}$ function to the matrix $\mathbf{P}$, equations (5) or (6).
Weight each column of $\mathbf{P}$ by using γ_{ij} values ($\mathbf{P}'$), equation (7).
Assign $\mathbf{t}_{\mathrm{Ts}}$ choosing the index of maximum value of each column in the decision vector $\mathbf{P}'$.

Fig. 3. Different steps of the ensemble algorithm.

Concerning time complexity, the proposed ensemble will be obviously more time consuming than the base classifier, since it will compute K different models instead of one. However, the models computed will be simpler than the original ones, as the classification problem joins neighbor classes.

IV. Experiments

Several benchmark datasets with different characteristics have been tested in order to validate the methodology proposed. Table I shows the characteristics of these datasets, where the number of patterns, attributes, classes and the class distribution (number of patterns per class) can be seen. These publicly available real ordinal classification datasets were extracted from benchmark repositories (UCI [30] and *mldata.org* [31], [32]). Also, some of the ordinal regression benchmark datasets (pyrim, machine, housing and abalone) provided by Chu et. al [12] were considered since they are widely used in the ordinal regression literature [4], [17]. These datasets do not originally represent ordinal classification tasks but regression ones. To turn regression into ordinal classification, the target variable is discretized into K different bins (representing classes, in this case K was assigned to 5

TABLE I
CHARACTERISTICS OF THE BENCHMARK DATASETS, ORDERED BY THE NUMBER OF CLASSES

Dataset	#Pat.	#Attr.	#Classes	Class distribution
squash-stored (SS)	52	51	3	(23, 21, 8)
squash-unstored (SU)	52	52	3	(24, 24, 4)
tae (TA)	151	54	3	(49, 50, 52)
newthyroid (NT)	215	5	3	(30, 150, 35)
car (CA)	1728	21	4	(1210, 384, 69, 65)
eucalyptus (EU)	736	91	5	(180, 107, 130, 214, 105)
pyrimx5 (P5)	74	27	5	(15, 15, 15, 15, 14)
machinex5 (M5)	209	7	5	(42, 42, 42, 42, 41)
housingx5 (H5)	506	14	5	(101, 101, 101, 101, 101)
abalonex5 (A5)	4177	11	5	(836, 836, 835, 835, 835)
automobile (AU)	205	71	6	(3, 22, 67, 54, 32, 27)
pyrimx10 (P10)	74	27	10	(8, 8, 8, 8, 7, 7, 7, 7, 7, 7)
machinex10 (M10)	209	7	10	(21, 21, 21, 21, 21, 21, 21, 21, 21, 20)
housingx10 (H10)	506	14	10	(51, 51, 51, 51, 51, 51, 50, 50, 50, 50)
abalonex10 (A10)	4177	11	10	(418, 418, 418, 418, 418, ... 418, 418, 417, 417, 417)

and 10), with equal frequency, as proposed in the previously mentioned works [4], [12], [17].

A. Methods compared

For an extensive analysis, several methods are compared. The proposed methodologies are applied using both SVM and KDA (and their adaptation to ordinal regression) as base methods. From now on, the methodologies are named as:

- Ensemble learning for ordinal regression using product combiner with SVM and KDA (EPS and EPK).
- Ensemble learning for ordinal regression using average combiner with SVM and KDA (EAS and EAK).
- Simple ensemble learning for ordinal regression with SVM and KDA (SS and SK).

These results have been compared with other state-of-the-art ordinal and nominal methods, such as:

1) Ordinal methods:

- Kernel discriminant learning for ordinal regression (KDO) [4] and support vector ordinal regression with implicit constraints (SVOI) [17], methods used as base classifiers in the ensemble proposals.
- Ordinal class classifier using the C4.5 as base classifier (OCC) [3] and ordinal class classifier with specific ordinal weights (OCCW) and SVM [7], both discussed in Section II-B since they are closely related to the proposal.
- Extreme learning machine for ordinal regression (ELMOR) [33] because the Extreme Learning Machine paradigm has demonstrated good scalability and generalization performance with a faster learning speed when compared to SVM [34].
- The POM algorithm [2] introduced in Section II-C.

2) Nominal methods:

- SVM classifier with one-vs-one methodology (SVM1) [35], and one-vs-all formulation (SVMA) [35]. These are the two main approaches for dealing with multiclass problems when using binary classifiers. Both are closely related to the proposal and it seems necessary to verify if they yield similar performances.
- SVM classifier using a probabilistic reformulation of the one-vs-all paradigm (SVMPA). In this case, the one-vs-all approach is reformulated to estimate probabilities, just like the proposal in this work. That is, after performing each binary classification, the probabilities for each hypothesis are calculated and later combined by using a combination function (the product, as it has been the one presenting the best results in this experimental section). The purpose of this comparison is to check if the possible improvement of the ELOR method compared to the standard 1VsAll approach is due to the probabilistic component of the proposal. Equally distributed probabilities are considered and a weighting probability method is not necessary, because all the classes receive the same attention.
- AdaBoostM1 using C4.5 as base classifier (AdaB.). This ensemble classifier is one of the most widely used in the machine learning literature, given its proven performance.

KDO and the proposed ensemble variants were implemented using Matlab, as well as the POM model available through the `mnrfit` function. The authors of SVOI provide a publicly available software[1], which was considered both for the standalone version and the proposed ensemble version. The well-know `libsvm` implementation[2] was considered for all the different versions of the SVM ensembles and for OCCW. The Matlab code for ELM[3] was adapted to implement ELMOR. Finally, the Weka[4] machine learning framework [36] provided the implementations for OCC and AdaB.

B. Evaluated measures

Several measures can be considered for evaluating ordinal classifiers. The most common ones in machine learning are the mean absolute error (MAE) and the mean zero-one error (MZE) [4], [17], [18], being $MZE = 1 - Acc$, where Acc is the accuracy or correct classification rate. However, as previously said, these measures are not the best option, for example, when measuring performance in the presence of class imbalances [37] and/or when the costs of different errors vary markedly. Because of that, this work makes use of different kind of measures to evaluate classifier performance.

The mean absolute error (MAE) is the average deviation in absolute value of the predicted class from the true class [37]: $MAE = \frac{1}{N}\sum_{i=1}^{N} e(\mathbf{x}_i)$, where $e(x_i) = |r(y_i) - r(y_i^*)|$ is the distance between the true and the predicted ranks ($r(y)$ being the rank for a given target y), and, then, MAE values range from 0 to $K-1$ (maximum deviation in number of ranks between two labels).

The average mean absolute error ($AMAE$) is the mean of the MAE across classes [37]: $AMAE = \frac{1}{K}\sum_{k=1}^{K} MAE_k = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k} e(\mathbf{x}_i)$, where $AMAE$ values range from 0 to $K-1$.

The maximum mean absolute error ($MMAE$) for all the classes is the MAE value considering only the patterns from

[1]http://www.gatsby.ucl.ac.uk/~chuwei/svor.htm
[2]http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[3]http://www.ntu.edu.sg/home/egbhuang/
[4]http://www.cs.waikato.ac.nz/ml/weka/

the class with the greatest distance between true labels and predicted ones: $MMAE = \max\{MAE_k; k \in \{1,\dots,K\}\}$, where MAE_k is the MAE value considering only the patterns from the k-th class and N_k is the number of pattern in this class. $MMAE$ values range from 0 to $K-1$. This measure was recently proposed [38] and its advantage is that a low $MMAE$ represents a low error for all independently considered classes.

The Kendall's τ_b is a statistic used to measure the association between two measured quantities. Specifically, it is a measure of rank correlation: $\tau_b = \frac{\sum c^*_{ij} c_{ij}}{\sqrt{\sum c^{*2}_{ij} \sum c^2_{ij}}}, i \in \{1,\dots,N\}, j \in \{1,\dots,N\}$, where c^*_{ij} is $+1$ if y^*_i is greater than (in the ordinal scale) y^*_j, 0 if y^*_i and y^*_j are the same, and -1 if y^*_i is lower than y^*_j, and the same for c_{ij}. τ_b values range from -1 (maximum disagreement between prediction and true label), to 0 (no correlation between them) and to 1 (maximum agreement). One important advantage of this correlation index is that it makes no assumption about the scale of the ranks.

The weighted Kappa (W_k) is a modified version of the Kappa statistic to allow different weights to different levels of aggregation between two variables: $W_\mathrm{k} = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}$, with $p_{o(w)} = \frac{1}{n}\sum_{i=1}^K \sum_{j=1}^K w_{ij} n_{ij}$, and $p_{e(w)} = \frac{1}{n^2}\sum_{i=1}^K \sum_{j=1}^K w_{ij} n_{i\cdot} n_{\cdot j}$, where n_{ij} is the number of times the patterns are predicted by the classifier to be in class j when they really are in class i, $n_{i\cdot} = \sum_{j=1}^K n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^K n_{ij}$ for $i, j \in \{1,\dots,K\}$. The weight $w_{ij} = |i-j|$ quantifies the degree of discrepancy between true (y_i) and predicted (y^*_j) categories, and W_k range from -1 to 1.

In this sense, different character measures are used. Firstly, the Acc measure, the most common for classification, reports, in terms of a ratio, how well the classifier works without making any distinction between the classes in the problem. Secondly, the standard MAE measure, well-known for ordinal regression problems, considers different misclassification errors. Also, two novel measures are used in order to prove whether the proposal achieves more balanced predictions when the number of patterns is very different for each class. The $AMAE$ metric reports how well all the classes are classified and the $MMAE$ gives information about the worst classified class. Finally, two different statistics are considered, in order to measure the association between prediction and true labelling.

C. Evaluation and model selection

Regarding the experimental setup, a holdout stratified technique was applied to divide the datasets 30 times, using 75% of the patterns for training and the remaining 25% for testing. For the regression datasets provided by Chu et. al [12] (pyrim, machine, housing and abalone), the number of random splits was 20 and the number of training and test patterns are the same as those presented in the corresponding works [12], [17]. The partitions were the same for all methods compared and one model was obtained and evaluated (in the test set), for each split. Finally, the results are taken as the mean and standard deviation of the measures over the 30 test sets.

The parameters of each algorithm are chosen using a nested validation with each of the training sets (k-fold method with $k = 5$) and the cross-validation criteria is the MAE since it can be considered the most common one in ordinal regression. The kernel selected for all the algorithms is the Gaussian one, $K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)$ where σ is the standard deviation.

For every tested kernel method (KDA and SVM methods), the kernel width was selected within these values $\{10^{-3}, 10^{-2}, \dots, 10^3\}$, as the cost parameter associated with SVM methods. The parameter u for avoiding singularity (for the methods based on KDA) was selected within $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, and the C parameter for the KDO was selected within the following ones $\{10^{-1}, 10^0, 10^1\}$.

D. Results

This section presents three different types of experiments. Firstly, a synthetic dataset is designed in order to show the advantages of the proposal graphically when comparing it with the one-vs-all standard formulation. Secondly, the results obtained are compared for the 15 datasets previously presented, with the 6 ensemble methodologies proposed and 10 state-of-the-art algorithms, using a set of 6 different selected measures. Finally, a different set of experiments with large-scale ordinal datasets is performed to analyze the potential scalability and interpretability of the proposed method.

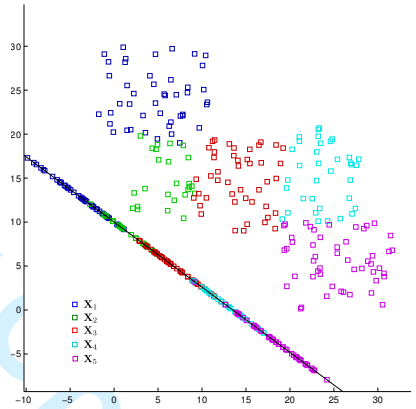

Fig. 4. Synthetic dataset and the optimal projection computed by linear discriminant analysis for ordinal regression [4].

1) Graphical representation of the proposal: In this subsection, a new synthetic dataset has been designed in order to show the main differences and advantages when comparing the ordinal version of the algorithms and the one-vs-all standard proposal. The graphical representation of the dataset can be seen in Figures 4 and 5.

Figure 4 shows the ordinal projection computed and the projected patterns for the dataset. Linear discriminant analysis for ordinal regression has been used for this (without using the kernel trick) to allow the representation of the results, due to the fact that the kernel trick would classify the dataset structure perfectly. Taking into account the final projection, it can be observed that classes 2, 3 and 4 are not very well-classified since they present some overlapping on the projection.

On the other hand, Figure 5 shows various projections computed and the patterns projected by the proposed procedure and the one-vs-all paradigm (also using linear discriminant analysis). The computed projections for the first and last classes

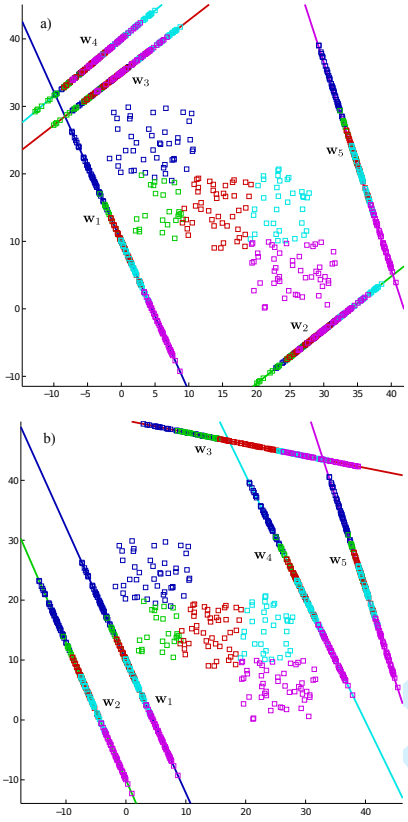

Fig. 5. Graphical representation of the different projections computed for the synthetic dataset using linear discriminant analysis: a) Projections computed by the one-vs-all formulation. b) Projections computed by the ordinal ensemble methodology proposed.

are seen to be the same as in the previous case, since the classification tasks are the same. But in this case, the projection for the rest of the classes allows better separation since some order among the classes is supposed. For example, $\mathbf{w}_3$ allows a clearer separation for the classes than the computed $\mathbf{w}_3$ in the one-vs-all formulation, where the classes $\{1, 2, 3\}$ are mixed in the projection. Each single class is seen to be well-classified in at least one model, and also the classes are ordered in the projection so that some information is implicit in the model.

2) Experimental results: The algorithms compared here have been run and optimized under the same conditions and using the same parameter cross-validation. First, the different ensemble proposals and their base algorithms are compared, and then the rest of the state-of-the-art methods are considered.

Table II shows the mean ranking for the proposals and the base methodologies for all 15 datasets, taking into account 6 different measures, which may help the reader to evaluate the value of the proposal. This table only considers the mean ranking (over all the datasets) obtained for each method and each metric. In this case (where 8 algorithms were compared), a ranking of 1 is assigned to the best method for a given dataset, and a ranking of 8 to the one which provides the worst performance. In this table and all the following ones, the best method is in bold face and the second one in italics. The mean ranking considering all the metrics has also been included in the table as a summary. In almost all cases, the ensemble

TABLE II
MEAN RANKINGS OF THE 15 DATASETS CONSIDERED FOR THE ENSEMBLE METHODOLOGIES PROPOSED AND THE BASE ALGORITHMS USED.

	Method							
Measure	KDO	EPK	EAK	SK	SVOI	EPS	EAS	SS
Acc	6.73	3.87	4.80	5.77	3.87	**2.07**	*3.50*	5.40
MAE	6.27	4.70	4.60	6.67	3.30	*2.67*	**2.60**	5.20
$AMAE$	6.27	4.70	4.37	6.27	*3.00*	**2.53**	3.33	5.53
$MMAE$	5.27	4.83	3.70	5.67	*3.53*	**3.40**	4.00	5.60
τ_b	5.47	3.97	4.57	7.00	3.27	**2.87**	*3.07*	5.80
W_k	5.73	3.57	5.23	6.87	*3.47*	**2.00**	3.60	5.53
Average	5.96	4.27	4.54	6.37	3.41	**2.59**	*3.35*	5.51

achieves better results than the initial algorithms. Specifically, it can be seen that the best results or the second best results for almost all the metrics tested are achieved by applying the EPS proposal. The complete tables of results showing the means and standard deviations for all benchmark datasets and metrics are not included in this work for the sake of simplicity and readability, but they can be found on a public webpage[5].

To quantify whether a statistical difference exists among the algorithms compared in Table II, a procedure is employed to compare multiple classifiers in multiple datasets [39]. First of all, a Friedman's non-parametric test with a significance level of $\alpha = 0.05$ has been carried out to determine the statistical significance of the differences in the mean ranking results for each measure selected. The test rejected the null-hypothesis that all algorithms perform similarly when $\alpha = 0.05$ for all the selected metrics, stating then that the differences in mean rankings of Acc, MAE, $AMAE$, $MMAE$, Kendall's τ_b and W_k are statistically significant. Specifically, the confidence interval for this number of datasets and algorithms is $C_0 = (0, F_{(\alpha=0.05)} = 2.10)$, and the corresponding F-value for each metric was $7.95 \notin C_0$, $9.30 \notin C_0$, $7.65 \notin C_0$, $2.47 \notin C_0$, $8.11 \notin C_0$ and $10.22 \notin C_0$ for Acc, MAE, $AMAE$, $MMAE$, Kendall's τ_b and W_k, respectively.

On the basis of this rejection, the Nemenyi post-hoc test is used to compare all classifiers to one another. This test considers that the performance of any two classifiers is deemed significantly different if their mean ranks differ by at least the critical difference (CD), which depends on the number of datasets and methods. 5% significance confidence was considered ($\alpha = 0.05$) to obtain this CD and the results can be observed in Figure 6, which shows CD diagrams as proposed in [39]. Each method is represented as a point on a ranking scale, corresponding to its mean ranking performance. CD segments are included to measure the separation needed between methods in order to assess statistical differences. Red lines group algorithms for which statistically significant, different mean ranking performance can not be assessed.

From the results of the statistical tests and from the tables, several conclusions can be drawn: firstly, one could notice by analysing mean rankings that the techniques based on SVMs present a clearly better performance than the ones based on KDA, and the ensemble procedure based on SVM usually outperforms the results obtained by the ensemble based on KDA, independently of the combiner or metric

[5]http://www.uco.es/grupos/ayrna/es/elor2013

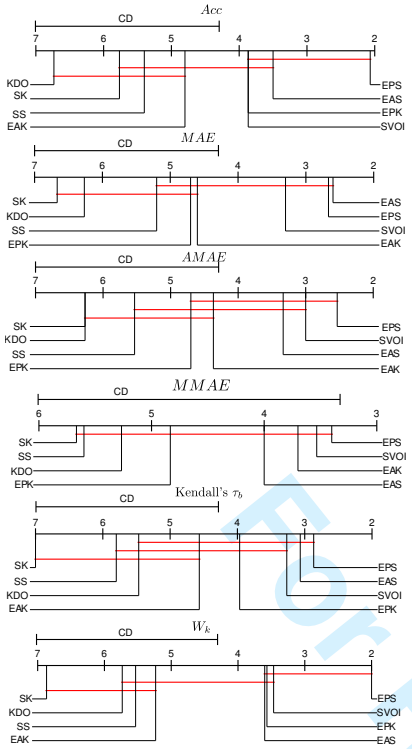

Fig. 6. Results and ranking of the Nemenyi statistical test for proposals and base methods.

used. Secondly, no significant differences can be observed by analysing different probability combiners, although the great majority of the results show better performance using the product combiner. Also, the methodology SELOR (SK and SS) can not be considered to be a good approach since its performance is worse than that of the base algorithms in many cases. Thus, it has been shown that considering all the probability information, performance can be significantly improved. Last but not least, the ensemble procedure seems to be a good approach to tackle ordinal regression since it leads to an improvement in the results obtained by several algorithms of the state-of-the-art (as the base classifiers used: KDO and SVOI) taking different measures into account. This can be easily seen by analysing the Nemenyi post-hoc figures.

To complete this section, a table similar to the previous one but containing the mean rankings for the rest of the state-of-the-art algorithms is shown in Table III. The EPS proposal is also included in this table, since, as stated before, it could be considered the proposal with the best performance. This table shows that the EPS procedure seems to be competitive for all measures (both ordinal and nominal), since it always obtains the best mean ranking. The second best method is OCCW, with the second position for all measures.

Table III shows that the EPS methodology is the best one in performance for all 6 metrics, improving the performances of 4 different ordinal classifiers and 4 nominal ones, and achieving a considerable balance between Acc, ordinal measures, those appropriate for imbalanced datasets, and correlation ones.

In this case, the non-parametric Friedman's test with a significance level of $\alpha = 0.05$ was also applied to the mean rankings for each measure. The test rejected the null-hypothesis that all algorithms perform similarly when $\alpha = 0.05$ for all the selected metrics, stating then that the differences in the mean ranking of Acc, MAE, $AMAE$, $MMAE$, Kendall's τ_b and W_k are statistically significant. Specifically, the confidence interval for this number of datasets and algorithms is $C_0 = (0, F_{(\alpha=0.05)} = 2.02)$, and the corresponding F-value for each metric was $12.33 \notin C_0$, $9.52 \notin C_0$, $7.08 \notin C_0$, $6.91 \notin C_0$, $11.24 \notin C_0$ and $11.63 \notin C_0$ for Acc, MAE, $AMAE$, $MMAE$, Kendall's τ_b and W_k, respectively.

TABLE III
MEAN RANKINGS OF THE 15 DATASETS FOR THE SELECTED ENSEMBLE METHODOLOGY AND OTHER STATE-OF-THE-ART METHODS.

	Method								
Measure	EPS	OCCW	OCC	ELMOR	POM	SVM1	SVMA	SVMPA	AdaB.
Acc	**2.03**	*2.60*	6.67	5.20	6.23	3.50	5.40	7.43	5.93
MAE	**1.63**	*3.07*	6.10	4.67	5.93	4.57	6.00	7.30	5.73
$AMAE$	**1.67**	*3.60*	6.20	4.80	5.40	4.70	5.90	7.00	5.73
$MMAE$	**1.67**	*4.00*	6.27	4.27	4.93	5.27	6.13	7.00	5.47
τ_b	**1.60**	*2.87*	6.60	4.73	5.20	4.53	6.00	7.33	6.13
W_k	**1.40**	*3.07*	6.73	5.27	5.27	4.27	5.73	7.33	5.93
Average	**1.67**	*3.20*	6.43	4.82	5.49	4.47	5.86	7.23	5.82

It is well-known that the Nemenyi approach comparing all classifiers to one another in a post-hoc test is not as sensitive as the approach comparing all classifiers to a given classifier (known as a control method) [39]. The Holm test performs this latter type of comparison, only considering the comparison between the control method and all the alternatives, and sequentially testing the hypotheses ranked by their significance. The ordered p-values will be denoted by $p_1 \leq p_2 \leq \ldots \leq p_{k-1}$, where k is the number of comparisons made. This step-down procedure compares p_i with a corrected version of the level of significance $\alpha/(k-1)$, starting with the most significant p-value (p_1). If p_i is below the corrected α, the null hypothesis is rejected and the next comparison is performed. When a certain null hypothesis can not be rejected, all the remaining ones are also retained. The results of this test (corrected α values and p-values) for all the measures are included in Table IV, where EPS is used as the control method.

This table shows that the EPS presents statistically significant differences for $\alpha = 0.05$ for almost all measures with respect to almost all methods, except for SVM1 (when using Acc) and OCCW (when using some of the metrics). No statistically significant differences could be assessed when comparing EPS to SVM1 for Acc, which is, in fact, a nominal method not designed to deal with ordinal problems. Furthermore, it can be seen that the proposal presents significant statistical differences for $\alpha = 0.05$ and the $MMAE$ metric with respect to the OCCW methodology, which could be considered the procedure most similar to the one designed in this work, and that the differences for $AMAE$ and W_k are also significant for $\alpha = 0.10$. In any case, it is important to remember that the mean rankings are always the best for EPS.

From these results, several conclusions can be drawn: firstly, as said before, it has been proven that an ordinal regression point of view is needed when dealing with some given order among categories, because, although a nominal algorithm may perform well when taking into account, for example, the

TABLE IV
RESULTS OF THE HOLM PROCEDURE USING EPS AS THE CONTROL METHOD WHEN COMPARED TO OTHER STATE-OF-THE-ART METHODS: CORRECTED α VALUES, COMPARED METHOD AND p-VALUES, ALL OF THEM ORDERED BY THE NUMBER OF COMPARISON (i).

			Acc		MAE		$AMAE$	
i	$\alpha^*_{0.05}$	$\alpha^*_{0.10}$	Method	p_i	Method	p_i	Method	p_i
1	0.0063	0.0125	SVMPA	0.0000$_\bullet$	SVMPA	0.0000$_\bullet$	SVMPA	0.0000$_\bullet$
2	0.0071	0.0143	OCC	0.0000$_\bullet$	OCC	0.0000$_\bullet$	OCC	0.0000$_\bullet$
3	0.0083	0.0167	POM	0.0000$_\bullet$	SVMA	0.0000$_\bullet$	SVMA	0.0000$_\bullet$
4	0.0100	0.0200	AdaB.	0.0001$_\bullet$	POM	0.0000$_\bullet$	AdaB.	0.0001$_\bullet$
5	0.0125	0.0250	SVMA	0.0008$_\bullet$	AdaB.	0.0000$_\bullet$	POM	0.0002$_\bullet$
6	0.0167	0.0333	ELMOR	0.0015$_\bullet$	ELMOR	0.0024$_\bullet$	ELMOR	0.0017$_\bullet$
7	0.0250	0.0500	SVM1	0.1425	SVM1	0.0034$_\bullet$	SVM1	0.0024$_\bullet$
8	0.0500	0.1000	OCCW	0.5709	OCCW	0.1518	OCCW	0.0532$_\circ$

			$MMAE$		τ_b		W_k	
i	$\alpha^*_{0.05}$	$\alpha^*_{0.10}$	Method	p_i	Method	p_i	Method	p_i
1	0.0063	0.0125	SVMPA	0.0000$_\bullet$	SVMPA	0.0000$_\bullet$	SVMPA	0.0000$_\bullet$
2	0.0071	0.0143	OCC	0.0000$_\bullet$	OCC	0.0000$_\bullet$	OCC	0.0000$_\bullet$
3	0.0083	0.0167	SVMA	0.0000$_\bullet$	AdaB.	0.0000$_\bullet$	AdaB.	0.0000$_\bullet$
4	0.0100	0.0200	AdaB.	0.0001$_\bullet$	SVMA	0.0000$_\bullet$	SVMA	0.0000$_\bullet$
5	0.0125	0.0250	SVM1	0.0003$_\bullet$	POM	0.0003$_\bullet$	POM	0.0001$_\bullet$
6	0.0167	0.0333	POM	0.0011$_\bullet$	ELMOR	0.0017$_\bullet$	ELMOR	0.0001$_\bullet$
7	0.0250	0.0500	ELMOR	0.0093$_\bullet$	SVM1	0.0034$_\bullet$	SVM1	0.0042$_\bullet$
8	0.0500	0.1000	OCCW	0.0196$_\bullet$	OCCW	0.2053	OCCW	0.0956$_\circ$

$\bullet$: Statistical difference with $\alpha = 0.05$
$\circ$: Statistical difference with $\alpha = 0.10$

measure of accuracy, it may fail when taking into account other ordinal measures. Secondly, as statistically significant differences exist for all the metrics selected when taking into account the different one-vs-all proposals (the nominal proposal for reformulating the SVM paradigm and the proposal in this work), ELOR seems to present clear advantages over the one-vs-all nominal paradigm, when tackling ordinal classification. Finally, it can be concluded that the combination of single classifiers, aiming at a more accurate classification decision at the expense of increased complexity, seems to be a good idea in this case, since it improves the performance of other state-of-the-art methodologies significantly.

3) Large-scale datasets and interpretability: Once the performance of the proposed method has been extensively validated making statistical comparisons to other state-of-the-art methodologies for different measures and datasets, there are some unanswered issues such as the scalability of the algorithm or its possible interpretability, which is the main aim of this subsection. However, these issues are more related to the choice of the base algorithm for the ensemble because it will obviously determine if the algorithm could be used with large-scale datasets or for model interpretability purposes. The complexity of the kernel methods previously used as base methodologies for the ensemble depend directly on the number of training patterns [4] and their interpretability is difficult. Because of this reason, a simpler and more interpretable method is used for the following experiments. This method does not present parameters to optimize and it is also designed for ordinal regression. It is a linear model, leading generally to a lower performance (see Table III of this paper or other studies in the ordinal classification literature [18], [19]). However, it provides us with a probabilistic output, a simpler model and a better interpretability. The method used is the POM algorithm [2] which was used for comparison purposes in the previous experimental subsection. Moreover, standard binary LR is used for the binary decompositions. This methodology, can be considered as interpretable in the sense that it could give us clues about the importance of each attribute for modelling the dependent variable.

For the experiments, two real ordinal datasets have been used. First, the Happiness dataset was extracted from the "European Social Survey"[6] considering year 2010 and 26 countries. It represents the complex problem of predicting the individual happiness by using certain characteristics, beliefs and life circumstances in a Likert scale (examples of some input variables are: the health of the person, if he or she has anyone to discuss personal matters, whether he or she takes part in social activities, etc). We selected 13 attributes and considered 5 classes. The dataset was composed of 41472 instances (missing values were removed for simplicity). For more information of this dataset see the webpage associated to this paper[5]. Secondly, the SpanishFleet dataset was obtained from the "Fleet Register On the Net" considering year 2012 and the whole Spanish fleet to predict the commitment to sustainability of the Spanish vessels, using a categorization of the overexploitation of the gears employed provided by the Food and Agriculture Organization of the United Nations. This dataset was composed of 10460 instances, 6 attributes and 10 classes. For more information of this dataset see [40].

Concerning the experiments on these datasets, the same aforementioned experimental design was used (i.e., 30 random repetitions of a stratified holdout, with 75% for training and 25% for the test set). To analyze the scalability of the algorithms, the complete time in seconds for executing each algorithm is also included in the results (note that the same machine architecture was used). The methods tested are: 1) the POM algorithm (which was previously presented) 2) ordinal class classifier using POM as base classifier (OCCP) and 3) ensemble learning for ordinal regression using product combiner and the POM algorithm as base classifier (EPP). We considered OCCP because it is a decomposition method which can be said to follow the same philosophy of EPP but using binary classifiers.

The results of these experiments can be seen in Table V. From these results, it can be seen that the proposed method outperforms in all the metrics the base classifier and, in most of the cases, the ordinal binary decomposition method (OCCP), thus providing more robust results. Furthermore, although both classification problems are complex because of the variable to predict, the obtained results are very promising (e.g., in MAE and $AMAE$). With regard to the execution time, the computational complexity of the methodology is affordable, even for large-scale problems. Furthermore, as it can be seen in the experiments (comparing the time obtained in both datasets), the time complexity of the algorithm depends to a greater extent on the number of classes (because it determines the number of decompositions to perform) rather than on the number of samples.

Concerning interpretability, the decomposition proposed provides us with additional information in the sense that one model for differentiating each class from the previous and following classes is computed. Therefore, instead of being pro-

[6]http://ess.nsd.uib.no/

TABLE V
MEAN TEST VALUES FOR THE DIFFERENT METHODS CONSIDERED.

Metrics	Happiness POM	Happiness OCCP	Happiness EPP
Acc	60.78 ± 0.15	$\mathit{63.44 \pm 0.26}$	$\mathbf{63.73 \pm 0.25}$
MAE	0.449 ± 0.002	$\mathit{0.402 \pm 0.003}$	$\mathbf{0.397 \pm 0.003}$
$AMAE$	1.259 ± 0.011	$\mathit{1.028 \pm 0.016}$	$\mathbf{1.002 \pm 0.014}$
$MMAE$	2.580 ± 0.051	$\mathit{1.953 \pm 0.081}$	$\mathbf{1.950 \pm 0.075}$
τ_b	0.232 ± 0.007	$\mathit{0.350 \pm 0.007}$	$\mathbf{0.375 \pm 0.006}$
W_k	0.088 ± 0.005	$\mathit{0.256 \pm 0.006}$	$\mathbf{0.293 \pm 0.005}$
Time	$\mathbf{23.41 \pm 4.34}$	$\mathit{32.49 \pm 0.50}$	49.00 ± 0.47
Metrics	SpanishFleet POM	SpanishFleet OCC(POM)	SpanishFleet EPP
Acc	83.33 ± 0.52	$\mathbf{86.62 \pm 0.39}$	$\mathit{85.87 \pm 0.28}$
MAE	0.443 ± 0.012	$\mathit{0.406 \pm 0.015}$	$\mathbf{0.388 \pm 0.010}$
$AMAE$	$\mathit{2.104 \pm 0.048}$	2.200 ± 0.117	$\mathbf{1.943 \pm 0.062}$
$MMAE$	6.880 ± 0.116	$\mathbf{5.996 \pm 0.370}$	$\mathit{6.594 \pm 0.151}$
τ_b	$\mathit{0.611 \pm 0.013}$	0.602 ± 0.021	$\mathbf{0.631 \pm 0.017}$
W_k	0.620 ± 0.012	$\mathit{0.665 \pm 0.013}$	$\mathbf{0.678 \pm 0.008}$
Time	$\mathbf{44.21 \pm 2.24}$	$\mathit{173.52 \pm 1.33}$	225.18 ± 2.11

TABLE VI
BEST SET OF MODELS $D_i, 1 \leq i \leq 5$, OBTAINED BY THE PROPOSED ORDINAL ENSEMBLE USING THE POM ALGORITHM AS BASE METHOD. THE MEANING OF EACH VARIABLE CAN BE FOUND IN THE WEBSITE ASSOCIATED TO THIS PAPER.

	D_1	D_2	D_3	D_4	D_5
$\mathbf{w}_1$	0.3172	0.1249	0.1065	0.0427	-0.1027
$\mathbf{w}_2$	0.0194	0.1485	0.1837	0.1829	0.1323
$\mathbf{w}_3$	0.2566	0.2196	0.1348	0.1175	0.0940
$\mathbf{w}_4$	1.0764	0.6465	0.5346	0.4028	0.1987
$\mathbf{w}_5$	0.1906	0.0614	0.0799	0.0551	-0.0239
$\mathbf{w}_6$	0.0873	0.2394	0.2218	0.1990	0.1711
$\mathbf{w}_7$	-0.2139	-0.1879	-0.2126	-0.2018	-0.0667
$\mathbf{w}_8$	-0.0257	0.0983	0.0774	0.0811	0.0954
$\mathbf{w}_9$	-0.0651	-0.1359	-0.1667	-0.1457	-0.0851
$\mathbf{w}_{10}$	-0.6461	-0.5314	-0.4974	-0.4229	-0.2605
$\mathbf{w}_{11}$	0.2036	0.1255	0.0849	0.0584	0.0083
$\mathbf{w}_{12}$	0.3355	0.2957	0.2535	0.1768	0.0819
$\mathbf{w}_{13}$	0.0358	-0.1183	-0.1664	-0.1968	-0.3089
b_1	-6.6354	-6.0191	-3.4551	-0.9498	2.5098
b_2	-	-3.6051	-1.0537	2.8159	-

TABLE VII
RANKING OF VARIABLES FOR THE DIFFERENT MODELS.

D_1	D_2	D_3	D_4	D_5
$\mathbf{x}_4$	$\mathbf{x}_4$	$\mathbf{x}_4$	$\mathbf{x}_{10}$	$\mathbf{x}_{13}$
$\mathbf{x}_{10}$	$\mathbf{x}_{10}$	$\mathbf{x}_{10}$	$\mathbf{x}_4$	$\mathbf{x}_{10}$
$\mathbf{x}_{12}$	$\mathbf{x}_{12}$	$\mathbf{x}_{12}$	$\mathbf{x}_7$	$\mathbf{x}_4$
$\mathbf{x}_1$	$\mathbf{x}_6$	$\mathbf{x}_6$	$\mathbf{x}_6$	$\mathbf{x}_6$
$\mathbf{x}_3$	$\mathbf{x}_3$	$\mathbf{x}_7$	$\mathbf{x}_{13}$	$\mathbf{x}_2$
$\mathbf{x}_7$	$\mathbf{x}_7$	$\mathbf{x}_2$	$\mathbf{x}_2$	$\mathbf{x}_1$
$\mathbf{x}_{11}$	$\mathbf{x}_2$	$\mathbf{x}_9$	$\mathbf{x}_{12}$	$\mathbf{x}_8$
$\mathbf{x}_5$	$\mathbf{x}_9$	$\mathbf{x}_{13}$	$\mathbf{x}_9$	$\mathbf{x}_3$
$\mathbf{x}_6$	$\mathbf{x}_{11}$	$\mathbf{x}_3$	$\mathbf{x}_3$	$\mathbf{x}_9$
$\mathbf{x}_9$	$\mathbf{x}_1$	$\mathbf{x}_1$	$\mathbf{x}_8$	$\mathbf{x}_{12}$
$\mathbf{x}_{13}$	$\mathbf{x}_{13}$	$\mathbf{x}_{11}$	$\mathbf{x}_{11}$	$\mathbf{x}_7$
$\mathbf{x}_8$	$\mathbf{x}_8$	$\mathbf{x}_5$	$\mathbf{x}_5$	$\mathbf{x}_5$
$\mathbf{x}_2$	$\mathbf{x}_5$	$\mathbf{x}_8$	$\mathbf{x}_1$	$\mathbf{x}_{11}$

vided with a model for tackling the whole learning problem, we obtain a model for discriminating each class and we could analyze independently the variables most determining.

To better visualize the interpretability of the model, let us analyze an example with the Happiness dataset. The best model (in this case the one performing better in terms of MAE for EPP) has been selected for the analysis. This model can be seen in Table VI. Note that both D_1 and D_5 are binary classifiers with a single threshold. The most important variables for modelling the labelling are the ones with higher $|\mathbf{w}_i|$ value, for example, it can be seen that $\mathbf{x}_4$ (satisfaction with present state of economy in country) presents a high impact on the variable to predict and so do $\mathbf{x}_{10}$ (the subjective health of the person). One should note that although the sign of $\mathbf{w}_i$ could also be used for an interpretability analysis, it could depend on the variable coding (in the case of the subjective health the variable is encoded from very good health to very bad health, thus this variable is negatively correlated with the label). Furthermore, it can be seen that variables important for different models are not so determining for others (analyze the case of $\mathbf{x}_1$, $\mathbf{x}_2$ or $\mathbf{x}_{13}$). Besides, as part of the model analysis, it can be said that having someone to discuss personal matters ($\mathbf{x}_7$) makes you happier (note that the "yes" have been encoded as 0 and "no" as 1).

As a final remark, if we order the variables taking into account their importance for each model (as said, the $|\mathbf{w}_i|$ value), it can be observed that some variables have almost no influence for discriminating certain classes (see Table VII). For example: being member of a group discriminated in your country or not ($\mathbf{x}_{11}$) is an influential variable for determining if you are extremely unhappy, but not for determining if you are extremely happy (it is at the last position). On the contrary, thinking that is important to help people and care for others well-being ($\mathbf{x}_{13}$) is indeed a determining variable for the happiest (it is at the first position).

V. CONCLUSIONS AND FUTURE WORK

The methodology here proposed is based on the computation of different classification tasks, by performing a relabelling process which takes ordinal data information into account. The relabelled data is then used for training the learning algorithm. In that sense, the proposal can be seen as a reformulation of the one-versus-all idea to tackle ordinal regression, as each single model is computed to differentiate each class from the remaining ones taking ordinal ranks into account. Threshold models are used as the base classifier because they are able to include the order information of these groups of classes and their natural projection capabilities facilitate the computation of probability estimations. For the prediction phase, two of the most widely studied combiners in the ensemble literature were used, the product and the average.

The proposal has been tested with 15 benchmark datasets and it has been found to be competitive when compared to the base classifiers and to other state-of-the-art methods. Statistical tests were applied to assess these conclusions. Additionally, the superiority of the proposal for the one-vs-all standard paradigm has been confirmed when dealing with ordinal regression. Although multiclass imbalance problems pose important difficulties for machine learning algorithms [41], this approach seems to achieve not only good global performance, but also good error rates for all classes independently, given the good $MMAE$ performance obtained.

Moreover, the proposal has been seen to be scalable (although this is an issue related to the base methodology, it has been seen to provide a reasonable time complexity compared to the base method) and interpretable (in the sense that the

most determining features for modelling each class can be extracted because it is based on a decomposition strategy).

Unlike discriminant analysis (where a normal distribution could be assumed), there is no guideline about the probability distribution to use when working with nonparametric approaches, such as SVMs. In fact, several studies have been performed in order to reformulate SVMs to allow probabilistic outputs [26], [27] making use of a maximum-likelihood estimator for adjusting the probability distribution to the projected patterns. This idea might be used as well in this work in order to compute fairer probabilities for the SVM methodologies.

Finally, the ensemble procedure could be tested with other ordinal base classifiers, also based on support vector machines or discriminant analysis such as those proposed in [8], [24].

ACKNOWLEDGMENT

This work has been partially subsidized by the TIN2011-22794 project of the Spanish Ministerial Commission of Science and Technology (MICYT), FEDER funds and the P2011-TIC-7508 project of the "Junta de Andalucía" (Spain).

REFERENCES

[1] M. Pérez-Ortiz, P. A. Gutiérrez, C. Hervás-Martínez, J. Briceño, and M. de la Mata, "An ensemble approach for ordinal threshold models applied to liver transplantation," in *Proceedings of the the 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 2795–2802.

[2] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society*, vol. 42, no. 2, pp. 109–142, 1980.

[3] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Proc. of the 12th Eur. Conf. on Machine Learning*, 2001, pp. 145–156.

[4] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, and W.-B. Li, "Kernel discriminant learning for ordinal regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 906–910, 2010.

[5] S. Kramer, G. Widmer, B. Pfahringer, and M. de Groeve, "Prediction of ordinal classes using regression trees," in *Proceedings of 12th International Syposium on Methodologies for Intelligent Systems (ISMIS 2000)*, ser. Lecture Notes in Computer Science, vol. 1932/2010, Charlotte, NC, USA, October 1114 2000, pp. 665–674.

[6] S. B. Kotsiantis and P. E. Pintelas, "A cost sensitive technique for ordinal classification problems," in *Proceedings of the Third Hellenic Conference on Artificial Intelligence (SETN2004)*, ser. Lecture Notes in Computer Science, vol. 3025/2004, Samos, Greece, May 5-8 2004, pp. 220–229.

[7] W. Waegeman and L. Boullart, "An ensemble of weighted support vector machines for ordinal regression," *International Journal of Computer Systems Science and Engineering*, vol. 3, no. 1, pp. 1–7, 2009.

[8] L. Li and H.-T. Lin, "Ordinal Regression by Extended Binary Classification," in *Advances in Neural Inform. Processing Syst. 19*, 2007.

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

[10] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, Cambridge, 2003, pp. 937–944.

[11] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed., ser. Monog. on Stat. and Applied Prob. Chapman & Hall/CRC, 1989.

[12] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *Journal of Machine Learning Research*, vol. 6, pp. 1019–1041, 2005.

[13] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[14] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2009.

[15] S. Mika, "Fisher discriminant analysis with kernels," Ph.D. dissertation, Berlin, Dec. 2002.

[16] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, 1st ed. Cambridge University, 2000.

[17] W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural Computation*, vol. 19, no. 3, pp. 792–815, March 2007.

[18] P. A. Gutiérrez, M. Pérez-Ortiz, F. Fernandez-Navarro, J. Sánchez-Monedero, and C. Hervás-Martínez, "An Experimental Study of Different Ordinal Regression Methods and Measures," in *7th International Conference on Hybrid Artificial Intelligence Systems (HAIS)*, ser. Lecture Notes in Computer Science, vol. 7209, 2012, pp. 296–307.

[19] J. Sánchez-Monedero, P. A. Gutiérrez, P. Tino, and C. Hervás-Martínez, "Exploitation of Pairwise Class Distances for Ordinal Classification," *Neural Computation*, vol. In press, 2013.

[20] S. Menard, *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications, 2009. [Online]. Available: http://books.google.es/books?id=KuRWdnoe4WUC

[21] B. Fang, Y. Y. Tang, Z. Shang, and B. Xu, "Generalized discriminant analysis: A matrix exponential approach," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 1, pp. 186–197, 2010.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *International Conference on Artificial Neural Networks*, 1999, pp. 97–102.

[24] A. Shashua and Levin, *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, 2003, vol. 15, ch. Ranking with large margin principle: Two approaches, pp. 937–944.

[25] R. Anand, K. Mehrotra, C. Mohan, and S. Ranka, "Efficient classification for multiclass problems using modular neural networks," *Neural Networks, IEEE Transactions on*, vol. 6, no. 1, pp. 117 –124, jan 1995.

[26] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*. MIT Press, 1999, pp. 61–74.

[27] V. Franc, A. Zien, and B. Schölkopf, "Support vector machines as probabilistic models," in *ICML*, 2011, pp. 665–672.

[28] D. J. Miller and L. Yan, "Ensemble classification by critic-driven combining," in *Proceedings of the Acoustics, Speech, and Signal Processing, IEEE International Conference - Volume 02*, ser. ICASSP '99. Washington, DC, USA: IEEE Computer Society, pp. 1029–1032.

[29] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[30] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[31] PASCAL, "Pascal (pattern analysis, statistical modelling and computational learning) machine learning benchmarks repository," 2011. [Online]. Available: http://mldata.org/

[32] D. S. Sonnenburg, "Machine Learning Data Set Repository," 2011. [Online]. Available: http://mldata.org/

[33] W.-Y. Deng, Q.-H. Zheng, S. Lian, L. Chen, and X. Wang, "Ordinal extreme learning machine," *Neurocomputation*, vol. 74, no. 1-3, pp. 447–456, Dec. 2010.

[34] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[35] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Transaction on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explor. Newsletter*, vol. 11, pp. 10–18, 2009.

[37] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," in *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications (ISDA 09)*, Pisa, Italy.

[38] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, and P. A. Gutiérrez, "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)*, nov 2011.

[39] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[40] M. Pérez-Ortiz, R. Colmenarejo, J. Fernández, and C. Hervás-Martínez, "Can machine learning techniques help to improve the common fisheries policy?" in *Proceedings of the International Work Conference on Artificial Neural Networks (IWANN)*. Springer, Heidelberg (In press), 2013, pp. 278–286.

[41] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 4, pp. 1119–1130, 2012.