Max-Ordinal Learning

Inês Domingues, *Student Member, IEEE*, and Jaime S. Cardoso, *Senior Member, IEEE*

***Abstract*— In predictive modeling tasks, knowledge about the training examples is neither fully complete nor totally incomplete. Unlike semisupervised learning, where one either has perfect knowledge about the label of the point or is completely ignorant about it, here we address a setting where, for each example, we only possess partial information about the label. Each example is described using two (or more) different feature sets or views, where neither are necessarily observed for a given example. If a single view is observed, then the class is only due to that feature set; if more views are present, the observed class label is the maximum of the values corresponding to the individual views. After formalizing this new learning concept, we propose two new learning methodologies that are adapted to this learning paradigm. We also compare their instantiation in experiments with different base models and with conventional methods. The experimental results made both on real and synthetic data sets verify the usefulness of the proposed approaches.**

***Index Terms*— Classification, incomplete knowledge, ordinal data, supervised learning.**

I. Introduction

Learning from examples is one of the most successful areas in machine learning. This research area encompasses two fundamental learning frameworks: supervised and unsupervised learning, with different assumptions on the uncertainty in the training data. Supervised learning attempts to learn a concept to correctly label unseen instances, where the training instances have known labels, and therefore the ambiguity is at its minimum. Unsupervised learning attempts to learn the structure of the underlying sources of instances, where the labels of the training instances are unknown, and therefore the ambiguity is at its maximum.

However, in many applications, the knowledge about the training examples is neither complete nor totally incomplete. In frameworks like semisupervised learning, we either have perfect knowledge about the label of the point or we are completely ignorant about it. Still, this is not the only variation of supervised learning for problems with incomplete knowledge with interest in practical applications. Frameworks where, for each example, we only possess partial information about the label are of potential interest.

In this brief, we introduce and study the novel concept of max-ordinal learning (MOL). We extend and explore a preliminary study [1] in various directions. First, the new concept is formally formulated instead of heuristically motivated. Second, a new learning methodology is proposed as a natural solution for the new learning paradigm. The learning method presented in the preliminary study is also framed under the same paradigm. A further development is the use of methods adapted for ordinal data instead of conventional methods for nominal data. The experiments reported at the end of this brief include a thorough testing on synthetic and real data, greatly extending the initial results. We also renamed the new learning concept from “Max-Coupled Learning” to “MOL” to emphasize the role of the order information in the learning methodology.

A. Relevance

Although relevant to many fields, this brief presented here was inspired by a breast cancer application. When radiologists examine mammograms, they look for specific abnormalities. A breast can have mass, calcifications, or both. Based on the level of suspicion, lesions can be placed into one of seven breast imaging reporting and data system (BIRADS) [2] categories: 1) for no findings; 2) for benign findings; 3) for probably benign findings; 4) for suspicious findings; 5) when there is a large probability of malignancy; and 6) for proven cancer.

Under the perspective of a supervised learning setting, the prediction of the malignancy of a case could be addressed as multiclass classification problem, where there is a natural order among the classes: we would be therefore in the presence of an ordinal data classification problem, with higher BIRADS values corresponding to higher probabilities of malignancy. However, when more than one finding is present in the mammogram, the overall BIRADS in the medical report corresponds to the finding with the highest BIRADS. This is the key observation that motivated our work.

II. Problem Statement

In statistical pattern recognition, it is usually assumed that a training set of labeled patterns is available where each pair $\{\mathbf{x}_i, y_i\} \in \Re^d \times \mathcal{Y}$ has been generated independently from an unknown distribution. The goal is to induce a classifier, i.e., a function from patterns to labels $f : \Re^d \to \mathcal{Y}$. In this brief, we will focus on the ordinal case of $\mathcal{Y} = \{y_1, \ldots, y_K\}$, where $y_1 \prec \cdots \prec y_K$ and $\prec$ is a linear order relation in $\mathcal{Y}$.

MOL generalizes this problem by making significantly weaker assumptions about the labeling information. The labeled patterns in the training set can be one of three types that are organized into three different subsets, as follows:

1) $\mathcal{S}_1 = \{\mathbf{x}_i, y_i = f(\mathbf{x}_i)\}_{i=1}^{N_1}$, where $\mathbf{x}_i \in \Re^{d_1}$, $y_i \in \mathcal{Y}$ and $f(\cdot)$ is unknown.
2) $\mathcal{S}_2 = \{\mathbf{z}_i, y_i = g(\mathbf{z}_i)\}_{i=1}^{N_2}$, where $\mathbf{z}_i \in \Re^{d_2}$, $y_i \in \mathcal{Y}$ and $g(\cdot)$ is unknown.
3) $\mathcal{S}_{12} = \{\mathbf{x}_i, \mathbf{z}_i, y_i\}_{i=1}^{N_{12}}$, where $\mathbf{x}_i \in \Re^{d_1}$ and $\mathbf{z}_i \in \Re^{d_2}$, $y_i \in \mathcal{Y}$. It is known that $y_i = \max(f(\mathbf{x}_i), g(\mathbf{z}_i))$ but both $f(\mathbf{x}_i)$ and $g(\mathbf{z}_i)$ are unobserved.

Intuitively, when both views $\mathbf{x}$ and $\mathbf{z}$ are present, the individual classification of each view is unobserved and only the maximum of both is known. An illustration is given in Fig. 1 (note that $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_{12}$ are disjoint sets). When $\mathcal{S}_{12}$ is empty, this results in the learning of two independent classifiers.

In the above mentioned breast cancer application, y_i corresponds to a known classification in one of the six BIRADS ordinal classes present in the medical report. The subsets are as follows: 1) S_1 corresponds to the cases where only mass was detected in the mammogram; 2) S_2 corresponds to the cases where only calcifications

Manuscript received May 8, 2013; revised August 16, 2013; accepted October 20, 2013. Date of publication November 5, 2013; date of current version June 10, 2014. This work was supported by the National Funds through the Fundação para a Ciência e Tecnologia under Projects PTDC/SAUENB/114951/2009 and SFRH/BD/70713/2010.

The authors are with INESC TEC and Department of Engineering, University of Porto, Porto 4200-465, Portugal (e-mail: icsd@inescporto.pt; jaime.cardoso@ieee.org).

Digital Object Identifier 10.1109/TNNLS.2013.2287381

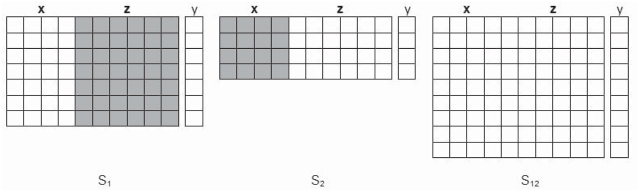

Fig. 1. Training set illustration. White represents observed and gray represents not present features.

were detected in the mammogram; and 3) S_{12} corresponds to the cases where both mass and calcifications were detected in the mammogram. $f(\mathbf{x}_i)$ corresponds to the BIRADS classification due to the presence of the mass only; and similarly for $g(\mathbf{z}_i)$. We stress that the report only includes the classification corresponding to the highest BIRADS.

The learning problem can be formulated as seeking $f(\cdot)$ and $g(\cdot)$ that minimize the expected loss over the distribution of observations, for a prespecified loss function. In general, the risk (expectation of the loss function) cannot be computed because the underlying distribution is unknown and functions $f(\cdot)$ and $g(\cdot)$ are selected based on the performance in the training set (empirical risk)

$$\{f^*, g^*\} = \arg\min_{f,g} \sum_{\mathbf{x}_i \in \mathcal{S}_1} \mathcal{L}(f(\mathbf{x}_i), y_i) + \sum_{\mathbf{z}_i \in \mathcal{S}_2} \mathcal{L}(g(\mathbf{z}_i), y_i) + \sum_{\mathbf{x}_i, \mathbf{z}_i \in \mathcal{S}_{12}} \mathcal{L}(\max(f(\mathbf{x}_i), g(\mathbf{z}_i)), y_i) \qquad (1)$$

where $\mathcal{L}$ is a loss function. A typical loss function for ordinal data is the mean absolute error (MAE). We emphasize that this problem formulation is valid only for ordinal data, since the maximum is not defined for nominal classes.

Although not fully explored in this manuscript, it is interesting to mention two special cases. The first is the extension of this model to multiple views, M (instead of just two). Each observation only includes a subset of the views and the label corresponding to the maximum of the individual views (the individual label of each view included is unknown). Another setting of interest is when all observations are from the third type; we always observe the two (or more) views and the label corresponds to the maximum of the unknown individual labels.

III. Related Work

In 2001, Frank and Hall [3] introduced a simple process which made it possible to explore information order in classification problems, using conventional binary classifiers. The problem is transformed from a K-class ordinal problem to $K - 1$ binary class problems. The main advantage of this scheme is that any binary classifier can be used as the building block.

In 2007, Cardoso presented the data replication method [4], a single binary classifier (SBC) reduction approach to solve multiclass problems via binary classification relying on a single, standard binary classifier. SBC reductions can be obtained by embedding the original problem in a higher dimensional space consisting of the original features, as well as one or more extension features. This embedding is implemented by replicating the training set points so that a copy of the original point is concatenated with each of the extension features' vectors. The binary labels of the replicated points are set to maintain a particular structure in the extended space. This construction results in an instance of an artificial binary problem, which is fed to a single binary learning algorithm. As in [3], any binary classifier can be used as the building block.

A kernel discriminant learning ordinal regression (KDLOR) method was proposed in 2010 [5]. KDLOR is an adaptation of the conventional linear discriminant analysis (LDA) method with a ranking constraint. The main goal is to find the optimal linear projection for classification (from which different classes can be well separated) while preserving the ordinal information of classes, i.e., the average projection of the samples from the higher rank classes should be larger than that of lower rank classes. The original LDA optimization problem is transformed and extended with a penalty term to account for the constraint in the projected means. To accommodate nonlinear problems, the algorithm is modified to incorporate the kernel trick.

More recently, a transductive ordinal learning (TOR) paradigm involving labeled and unlabeled data for learning ordinal decision functions was introduced [6]. A label swapping scheme for multiple ordinal class transduction was also proposed. Numerical results show that this transductive approach achieves significant accuracy improvements in terms of mean zero one and absolute errors.

In learning methodologies with incomplete knowledge, perhaps the most similar methodology to the one addressed here is multiple instance learning (MIL) [7]. The basic idea of MIL is that, during training, examples are presented in sets (often called "bags"), and labels are provided for the bags rather than for the individual instances. If a bag is labeled positive, it is assumed to contain at least one positive instance, otherwise the bag is negative. Note that this paradigm is for binary settings only and that all the observations in the bag come from the same "view" (feature set). Felzenszwalb *et al.* [8] use a latent variable formulation of the above mentioned MIL SVM [7] to train models using partially labeled data. Once again, this formulation only applies to the binary case. Furthermore, the classification function is the maximum of linear functions (in our approaches the functions do not need to be linear).

Techniques for semisupervised learning are common nowadays. Co-training and multiview models (assume that there are multiple, different learners trained on the same labeled data, and these learners agree on the unlabeled data) are representative examples. For example, in tri-training [9] the labeled data are split in three sets and a classifier is trained in each set. If two of them agree on the classification of an unlabeled point, the classification is used to "teach" the other classifier.

Note that some of the existing models cannot be applied to the present learning problem since they require all classifiers trained on the same data set (or on examples from the same population, with the same dimensions and features); other models do not take advantage of the order information in the classes to improve the generalization performance. Intuitively, they do not make the best usage of the information available in MOL for the learning process.

IV. Learning Max Ordinal Relations

Since we are exploring a new learning concept, one option is to adapt existing types of models (e.g., neural networks, support vector machines, etc.) to the new objective function. However, it would be interesting to accommodate this formulation under the ordinal class problem. This would allow the use of mature and optimized algorithms, developed for this well-established problem. We therefore discuss two alternative iterative processes that have, at the core, a base classifier for multiclass classification problems, which is not necessarily ordinal.

The proposed methodology makes use of a base classifier for each view. In general, the base classifiers can be from different types and they can be adapted for the data in the corresponding view (e.g., a SVM for the first view and a decision tree for the second). The methodologies to be presented do not make any assumptions in this respect. Regarding the ordinal nature of the data, the scenario is different. Both methodologies make use of the order in the splitting of the training data set in two, according to the predictions on each view. Inside the framework, the base classifier may or may not take advantage of the order information. We expect that classifiers that do make use of the order nature of the classes (including "more knowledge" in the learning process) achieve a better performance.

A. MOL Local Approximation Algorithm

Each base classifier is initialized by training it with all data from both the corresponding subset $\mathcal{S}_i$ and the subset $\mathcal{S}_{12}$. In the initialization, the labels in the subset $\mathcal{S}_{12}$ for each base classifier are assumed to correspond to the observed labels. For all subsequent iterations, consider the objective function in (1) rewritten as

$$\{f^*, g^*\} = \arg\min_{f,g} \sum_{\mathbf{x}_i \in \mathcal{S}_1} \mathcal{L}(f(\mathbf{x}_i), y_i) + \sum_{\mathbf{z}_i \in \mathcal{S}_2} \mathcal{L}(g(\mathbf{z}_i), y_i) + \sum_{\substack{\mathbf{x}_i, \mathbf{z}_i \in \mathcal{S}_{12} \\ f(\mathbf{x}_i) > g(\mathbf{z}_i)}} \mathcal{L}(f(\mathbf{x}_i), y_i) + \sum_{\substack{\mathbf{x}_i, \mathbf{z}_i \in \mathcal{S}_{12} \\ f(\mathbf{x}_i) < g(\mathbf{z}_i)}} \mathcal{L}(g(\mathbf{z}_i), y_i) \quad (2)$$

where the last term in (1) has been split in two, according to which of the views predicts the highest value.

Under a local approximation (LA) assumption that, in the next iteration, the order relation between the individual predictions $f(\cdot)$ and $g(\cdot)$ on the observations is kept, we optimize the current hypothesis (we use the terms classifier and hypothesis interchangeably) by retraining the model f on the set of observations $\mathcal{S}_1 \cup \{\mathbf{x}_i, \mathbf{z}_i : \mathbf{x}_i, \mathbf{z}_i \in \mathcal{S}_{12} \wedge f(\mathbf{x}_i) > g(\mathbf{z}_i)\}$ and by retraining the model g on the set of observations $\mathcal{S}_2 \cup \{\mathbf{x}_i, \mathbf{z}_i : \mathbf{x}_i, \mathbf{z}_i \in \mathcal{S}_{12} \wedge f(\mathbf{x}_i) < g(\mathbf{z}_i)\}$. Although the points in $\mathcal{S}_{12}$ with $f(\mathbf{x}_i) = g(\mathbf{z}_i)$ are not explicitly addressed in the above description, they are randomly split between the data sets used to update the two models. Furthermore, note that with this approach, in each iteration, each point is used when updating one of the models but not in both.

We stress that with MOL.LA any multiclass method without modifications can be selected for the base classifier. Naturally, we expect that base classifiers adapted for ordinal data achieve better performance than conventional classifiers for nominal data. It is also interesting to note that MOL.LA corresponds to the batch method proposed in our preliminary work [1], although it did not have a very strong theoretical support then and it was only instantiated with base classifiers for nominal data.

The adaptations of the MOL.LA algorithm for the two generalizations considered at the end of Section II are simple. In the extension of the model to M views, an observation can include any subset of views, corresponding to $2^M - 1$ different combinations. The base models are initialized as before, using the subset of observations containing the corresponding view and assuming that the label is due to that view. The iterative process also remains the same, where each model is retrained with the subset of observations where the model prediction is maximal. In the extreme case where every observation includes all views, the process remains the same as before.

B. MOL Coordinate Descent Algorithm

An alternative approach is to consider a coordinate descent (CD) methodology. The base classifiers are initialized as before but now, in each iteration, we perform two steps:

1) in the first step, we consider the hypothesis $g(\cdot)$ fixed and optimize the objective function over $f(\cdot)$;
2) in the second step, we consider the hypothesis $f(\cdot)$ fixed and optimize the objective function over $g(\cdot)$.

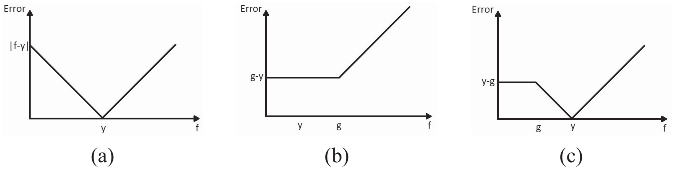

Fig. 2. MAE functions when g is fixed. (a) Subset $\mathcal{S}_1$. (b) Subset $\mathcal{S}_{12}$ and $g > y$. (c) Subset $\mathcal{S}_{12}$ and $g < y$ (similarly for the case when f is fixed).

In the first step, since $g(\cdot)$ is considered fixed, the optimization of (1) is equivalent to

$$f^* = \arg\min_{f} \sum_{\mathbf{x}_i \in \mathcal{S}_1} \mathcal{L}(f(\mathbf{x}_i), y_i) + \sum_{\mathbf{x}_i, \mathbf{z}_i \in \mathcal{S}_{12}} \mathcal{L}(\max(f(\mathbf{x}_i), g(\mathbf{z}_i)), y_i). \quad (3)$$

Splitting the last term in two gives

$$f^* = \arg\min_{f} \sum_{\mathbf{x}_i \in \mathcal{S}_1} \mathcal{L}(f(\mathbf{x}_i), y_i) + \sum_{\substack{\mathbf{x}_i, \mathbf{z}_i \in \mathcal{S}_{12} \\ f(\mathbf{x}_i) > g(\mathbf{z}_i)}} \mathcal{L}(f(\mathbf{x}_i), y_i) + \sum_{\substack{\mathbf{x}_i, \mathbf{z}_i \in \mathcal{S}_{12} \\ f(\mathbf{x}_i) < g(\mathbf{z}_i)}} \mathcal{L}(g(\mathbf{z}_i), y_i). \quad (4)$$

Since y_i and $g(\mathbf{z}_i)$ are assumed known and fixed, predictions by the hypothesis f above $g(\mathbf{z}_i)$ are penalized according to the adopted loss function. Predictions by the hypothesis f below $g(\mathbf{z}_i)$ are penalized by $\mathcal{L}(g(\mathbf{z}_i), y_i)$, which is independent of $f(\mathbf{x}_i)$. Considering MAE as the loss function for illustration purposes, an error in an observation in $\mathcal{S}_1$ is penalized as depicted in Fig. 2(a). Note that, although $f(\cdot)$ only assumes values in a finite set, a continuous representation was adopted in Fig. 2 for better visualization. The loss in an observation in $\mathcal{S}_{12}$ depends on the relative values of y and $g(\mathbf{z})$, as represented in Fig. 2(b) and (c). Note that with this approach, in each iteration, each model receives all points for training, although with different costs.

It is important to emphasize that this learning process incorporates two notions of distance or error. First, the base classifier is internally optimizing some notion of error, typically defined in the continuous domain. For instance, when instantiated with support vector machines, the internal loss is given by the hinge loss function. Externally, the framework is using a notion of loss defined in the space of the categorical values, $\mathcal{Y} = \{y_1, \ldots, y_K\}$, where $y_1 \prec \cdots \prec y_K$ and $\prec$ is a linear order relation in $\mathcal{Y}$. We use MAE as an example, but other options include using mean square error (MSE), average MAE (AMAE) or the ordinal classification index (OCI) [10].

The adaptations of the MOL.CD algorithm for the two generalizations considered at the end of Section II are also simple. The extension to M views is accomplished by fixing all but one of the M models at a time. Therefore, in each iteration, a total of M steps is performed.

C. Summation

The MOL.CD is a kind of CD in the space of the models. Like all gradient or CD methods, it can stay trapped in locally optimal solutions if the performance surface in the space of models is complex. In MOL.LA, the focus is on the partitioning of training instances into two subsets. Arguably, the most difficult part of the learning in MOL is understanding for which subset of points in S_{12} the label is due to the first view (subset $S_{12}^{(1)}$) and for which subset of points in S_{12} the label is due to the second view (subset $S_{12}^{(2)}$).

When this is known, the learning problem is now equivalent to the training of two independent classifiers, one in $S_1 \cup$ (subset $S_{12}^{(1)}$), the other in $S_1 \cup$ (subset $S_{12}^{(2)}$) (or semisupervised approaches, which likely provide better solutions). With MOL.LA, one tries to travel the space of partitions, choosing the next partition to be evaluated based on the predictions obtained by the classifiers trained in the current partition.

Both MOL.LA and MOL.CD are iterative methods that try to decrease the loss at each iteration. However, like many methods of this kind, the loss is not guaranteed to decrease monotonously. Moreover, for certain combinations of the loss function, base classifier, and data set the loss can fluctuate and the methods may not converge. Also, as is typical in these methods, they are either run for a prespecified number of iterations, or until there is no significant change in the models, or until the loss is below a prespecified quality value.

The prediction stage is common to both frameworks. Note that the output of the learning process for both frameworks is a set of classifiers, one per view, able to make predictions when receiving as input the attributes of the corresponding view.

When in the presence of a test instance, possibly only a subset of the views is present. The predicted output for the test instance will be the maximum of the individual predictions for each of the views that are present in the instance.

V. Experimental Validation

The data sets included in the experimental validation comprise real and synthetic data, with different levels of difficulty. While in data set $H_{2,5}H_{2,5}$ both views have the same distribution, in data set $C_{2,5}H_{3,5}$ the distribution of each view is different. A third data set, $H_{4,10}H_{4,10}$, was included to study the influence of the number of classes involved. Finally, data set $H_{3,5}C_{2,5}H_{3,5}$ was included to exemplify the three view case. A detailed description on how each data set was generated is provided in the following paragraphs.

Data Set $H_{2,5}H_{2,5}$: Example points $\mathbf{x} = (x_1, x_2)^t$ (dim = 2) were randomly generated in the unit square $[0, 1] \times [0, 1] \in \Re^2$ according to the uniform distribution [4]. Each point was assigned a class y from the set $\{1, 2, 3, 4, 5\}$, according to

$$y = \min_r \left\{ r : b_{r-1} < 10^{\dim -1} \prod_{i=1}^{\dim} (x_i - 0.5) + \epsilon < b_r \right\}$$
$$(b_0, b_1, b_2, b_3, b_4, b_5) = (-\infty, -1, -0.1, 0.25, 1, +\infty) \quad (5)$$

where $\epsilon \sim N(0, 0.125^2)$ simulates the possible existence of error in the assignment of the true class on $\mathbf{x}$.

The $\mathbf{z}$-view of the data was similarly generated. We generated N_1 samples for the $\mathbf{x}$-view only, N_2 samples for the $\mathbf{z}$-view only, and N_{12} samples for both views, where examples from both views were concatenated in a $\Re^4$-feature, keeping only the maximum of the corresponding labels for the observed output of the example. For an illustration of the individual views in this data set, see [4].

Data Set $C_{2,5}H_{3,5}$: The $\mathbf{x}$-view example points $\mathbf{x} = (x_1, x_2)^t$ were once again randomly generated in the unit square $[0, 1] \times [0, 1] \in \Re^2$ according to the uniform distribution. However, class y was assigned to each point according to the radius of the point, $\left\lceil \sqrt{x_1^2 + x_2^2} + \epsilon \right\rceil$ and using the limits

$$\left(0, \frac{1}{5}\frac{\sqrt{2}}{2}, \frac{2}{5}\frac{\sqrt{2}}{2}, \frac{3}{5}\frac{\sqrt{2}}{2}, \frac{4}{5}\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \quad (6)$$

where ϵ is as before. The $\mathbf{z}$-view was generated in a similar way to data set $H_{2,5}H_{2,5}$ but now in a 3-D space (dim = 3) according to (5), with the same b_i's and noise distribution.

Data Set $H_{4,10}H_{4,10}$: Data set $H_{4,10}H_{4,10}$ was built in a similar way to data set $H_{2,5}H_{2,5}$. However, each of the two views has 4-D (dim = 4) and 10 classes were generated by replacing 5 with

$$(b_0, b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, b_{10})$$
$$= (-\infty, -5, -2.5, -1, -0.4, 0.1, 0.5, 1.1, 3, 6, +\infty). \quad (7)$$

For an illustration of the views in this data set, see [4].

Data Set $H_{3,5}C_{2,5}H_{3,5}$: This data set has three views. The first and last views were generated as described for data set $H_{2,5}H_{2,5}$ and the middle view was generated with the same methodology as for the first view of data set $C_{2,5}H_{3,5}$.

Data Set INbreast: Mammograms from the INBreast database [11] and their respective annotations were used to create the real data sets. For the masses, a total of 15 features were extracted: anisotropy, area ratio, area, compactness, contor roughness, eccentricity, entropy of the intensity distribution, Haussdorf fractal dimension, inertial momentum, local concavity, mean of the intensity, roundness metric, solidity, standard deviation of the normalized radial length, and wavelet transform [12]. For calcifications, 10 features were used: average size, brightness, compactness, contrast, density, diffuseness, eccentricity, number of calcified spots, number of clusters, and the standard deviation of distances from the cluster centroid [11], [13].

Mammography comprehends the recording of two sights for each breast: the craniocaudal (CC) sight, which is a top to bottom sight, and a mediolateral oblique (MLO) sight, which is a side sight. Thus, two types of data sets were built. In data set “individual,” each sight was analyzed individually while in data set “combined” the two sights of each breast, MLO and CC, were classified together. Specifically, in the “combined” data set, we have a total of four views: masses of MLO sight, calcifications of MLO sight, masses of CC sight, and calcifications of CC sight.

Methodology: We randomly split the generated synthetic data sets into training and test sets. The total number of generated points for each data set was 1000 (except for data set $H_{4,10}H_{4,10}$ where, due to the higher number of classes, 2000 points were generated). 50% of all of the data used for training and the remaining 50% for testing. To study the effect of varying the proportion in the data with a single-view, we considered two possibilities: 10% and 40% of all of the data had information from a single-view (within this, the cases were equally divided by the two views). The splitting of the data was repeated 40 times to obtain more stable results for performance estimation.

The real data set was also divided into two nonoverlapping sets: 75% of the data was randomly selected for training and the remaining 25% was used for testing. As before, the splitting of the data was repeated 40 times. All features were normalized to have zero mean and unit variance.

As previously stated, and since the data is ordinal, we adopted MAE as a measure of performance. Each model parameterization was optimized by two-fold cross validation inside the training set. The nonordinal extension from binary to multiclass was conducted with the one-against-one method. The ordinal methods (already described in Section III) used were KDLOR, Frank and Hall, and data replication. With the exception of KDLOR, which is an extension of the LDA, all models were instantiated with SVMs. The MOL.CD framework was not instantiated in KDLOR and the nominal methods, since it is not clear how to incorporate misclassification costs into these base models. With the data replication and the Frank and Hall method, the misclassification cost of each observation is managed by controlling the presence of the observation in each data replica or in the training of each individual classifier.

TABLE I
MAE FOR THE SYNTHETIC AND REAL DATA SETS

	Dataset	Standard Model	Tri Training	MOL.LA				MOL.CD	
				Non-ordinal	KDLOR	Frank&Hall	Data Replication	Frank&Hall	Data Replication
10%	$H_{2,5}H_{2,5}$	0.372	0.908	0.206	0.186	0.196	0.191	0.193	0.189
	$C_{2,5}H_{3,5}$	0.599	0.953	0.297	0.220	0.225	0.229	0.193	0.166
	$H_{4,10}H_{4,10}$	2.267	2.316	1.233	0.601	0.565	0.571	0.581	0.592
	$H_{3,5}C_{2,5}H_{3,5}$	0.738	1.237	1.023	0.740	0.769	0.775	0.701	0.721
40%	$H_{2,5}H_{2,5}$	0.205	0.366	0.193	0.183	0.188	0.183	0.185	0.181
	$C_{2,5}H_{3,5}$	0.261	0.633	0.252	0.254	0.199	0.208	0.179	0.163
	$H_{4,10}H_{4,10}$	0.908	2.085	1.015	0.590	0.545	0.563	0.550	0.570
	$H_{3,5}C_{2,5}H_{3,5}$	0.246	1.082	0.555	0.401	0.398	0.367	0.400	0.393
individual	INbreast	0.372	0.413	0.282	0.576	0.270	0.268	0.257	0.259
combined	INbreast	0.227	0.401	0.216	0.577	0.210	0.255	0.196	0.218

For comparison purposes, both a standard multiclass method [1] and tri-training [9] were included in the experiments. In the standard multiclass, for the two-view example, the subset $\mathcal{S}_{12}$ is ignored while two models are derived, one for each view. In the test phase, the maximum predicted by the two classifiers is taken as the final prediction. A similar technique is used for the cases with more views. Four standard multiclass models were tested, one-against-one SVMs, KDLOR, Frank and Hall, and data replication. Only the best results (namely data replication) are presented.

Throughout this paper, we speak of two results as being "significantly different" if the difference is statistically significant at the 1% level according to a paired two-sided t-test, where each pair of data points consists of the estimates obtained in one of the 40 runs of the two learning schemes being compared.

A. Results

Table I summarizes the results (MAE) for the real and synthetic data sets.

The statistical significance analysis yielded the following results.

1) In all experiments, there was at least one instantiation with a base ordinal classifier of MOL.LA or MOL.CD that was statistically better than the two conventional methods and the nonordinal instantiation of MOL.LA.
2) In eight experiments (out of 10), tri-training was statistically worse than all the other models; in the other two experiments, it was statistically worse than all except MOL.LA instantiated with KDLOR.
3) In six experiments, Standard Model was statistically worse than all the MOL.LA and MOL.CD instantiations; in all the eight experiments with synthetic data, standard model was statistically worse than all the MOL.LA and MOL.CD instantiations with ordinal methods; in the two experiments with INbreast, it was statistically worse than all except MOL.LA instantiated with KDLOR.
4) Frank and Hall and data replication in MOL.CD were statistically better than the corresponding MOL.LA instantiation in four experiments; in the other six experiments there was no statistical difference.
5) MOL.LA with KDLOR was statistically worst than the corresponding MOL.LA instantiation with Frank and Hall (and the data replication) in three experiments; in the other seven experiments there was no statistical difference.
6) There is no method simultaneously statistically better than all the Frank and Hall and data replication instantiations, for any data set.

Several aspects are worth noting. A first observation is that both MOL.LA and MOL.CD techniques are better than standard methods. A legitimate conclusion is that, in a specific application scenario, it is enough to test and compare MOL.LA and MOL.CD instantiated with ordinal methods since there is always an instantiation superior to the other models. This conclusion reinforces our initial results (without significance analysis) on the advantage of models specific for this new learning problem [1]. Moreover, the results also support the use of models for ordinal data inside the proposed framework.

The recently proposed KDLOR behaves worse than both Frank and Hall and data replication. We point out that KDLOR is based on LDA while the Frank and Hall and data replication use SVMs. A formal study comparing KDLOR with the Frank and Hall and the data replication ideas by mapping the last two onto LDA is made elsewhere [14].

When comparing the two methodologies proposed in this brief, MOL.LA and MOL.CD, we observe that MOL.CD behaves better for some data sets, in particular for the real data set and for the synthetic data set $C_{2,5}H_{3,5}$. This performance advantage is counterbalanced by the increase in the time to design the models. Note that while with MOL.LA each observation is used to update one and only one of the models; with MOL.CD every observation is used by every model (albeit with different costs), which allows MOL.CD to do a better usage of the data.

For synthetic data, the increase in the number of views has a negative effect on accuracy. This is an expected outcome since we possess less information. For example, in S_{123} the available label corresponds to the maximum of three views whereas before it was the maximum of only two views.

In the experiments with real data, the inclusion of the two sights leads to an improvement in the accuracy. While this seems to contradict the results in the synthetic data set, in the real data set the views are correlated (which breaks the initial assumption). Recall that the four views are: 1) masses of MLO sight; 2) calcifications of MLO sight; 3) masses of CC sight; and 4) calcifications of CC sight.[1] This means that 1) and 3) correspond to the same location in the breast and thus they share the same label [similarly for 2) and 4)]. In this way, the label for each sample is the maximum of only two labels and not of four individual labels.

Concerning the performance, we first to point out that all the algorithms were implemented in MATLAB and no attempt was made to optimize the running times. Having this in mind, the fastest algorithms are the standard ones (tri-training and standard model), followed by MOL.LA; MOL.CD was the slowest methodology. Moreover, data replication implementation is slower than the corresponding Frank and Hall. When using these methodologies in practice, the trade-off between accuracy and performance must be considered.

[1] For each sample, each view may or may not be present.

VI. Conclusion

The typical learning settings already studied in the literature are not necessarily the most interesting for practical applications, since they may not represent well the information that is available. In this brief, we present the MOL paradigm, in between classification and semisupervised classification. For every observation, we do possess some information about the label. However, in a subset of the examples, the knowledge is incomplete. This corresponds to the worst-case classification of the individual views of the example.

A formalization of the MOL paradigm led to two new learning schemes. An evaluation of both synthetic and real data sets showed that the methodologies developed give better results. Both the use of ordinal schemes and the adaptation of the training itself to the max paradigm make the classifiers more suitable for the MOL problem. The "blind" use of traditional classifiers can attain sub-optimal results. The experiments here conducted underline the importance of including prior knowledge when designing a classifier.

One direction to continue this brief is on the use of reduction techniques (RT). Typically, RT can be obtained by embedding the original problem in a higher dimensional space; this embedding is implemented by replicating the training set points so that a copy of an original point is concatenated with extension feature vectors. The labels of the replicated points are set to maintain a particular structure in the extended space. This construction results in an artificially simpler problem, which is fed to a simpler learning algorithm. This idea has been applied to solve the MIL problem [15]–[17]. By proceeding in a similar way, we may be able to solve MOL using traditional ordinal learning methods. Moreover, reduction techniques have been used to solve multiclass classification problems [18], including the ordinal setting [4] with a SBC. Therefore, in the end, a single binary classifier could be used to address MOL.

References

[1] J. S. Cardoso and I. Domingues, "Max-coupled learning: Application to breast cancer," in *Proc. Int. Conf. Mach. Learn. Appl.*, 2011, pp. 13–18.

[2] C. J. Orsi, "The american college of radiology mammography lexicon: An initial attempt to standardize terminology," *Amer. J. Roentgenol.*, vol. 166, no. 4, pp. 779–780, 1996.

[3] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Proc. Eur. Conf. Mach. Learn.*, 2001, pp. 145–156.

[4] J. S. Cardoso and J. F. P. D. Costa, "Learning to classify ordinal data: The data replication method," *J. Mach. Learn. Res.*, vol. 8, pp. 1393–1429, Jul. 2007.

[5] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, and W.-B. Li, "Kernel discriminant learning for ordinal regression," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 906–910, Jun. 2010.

[6] C.-W. Seah, I. W. Tsang, and Y.-S. Ong, "Transductive ordinal regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1074–1086, Sep. 2012.

[7] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 561–568.

[8] P. F. Felzenszwalb, R. B. Girshick, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[9] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005.

[10] J. S. Cardoso and R. Sousa, "Measuring the performance of ordinal classification," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, no. 8, pp. 1173–1195, 2011.

[11] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Towards a full field digital mammographic database," *Academic Radiol.*, vol. 19, no. 2, pp. 236–248, 2011.

[12] D. Cascio, F. Fauci, R. Magro, G. Raso, R. Bellotti, F. De Carlo, *et al.*, "Mammogram segmentation by contour searching and mass lesions classification with neural network," *IEEE Trans. Nuclear Sci.*, vol. 53, no. 5, pp. 2827–2833, Oct. 2006.

[13] C.-H. Wei, Y. Li, and P. J. Huang, "Mammogram retrieval through machine learning within BI-RADS standards," *J. Biomed. Inf.*, vol. 44, no. 4, pp. 607–614, 2011.

[14] J. S. Cardoso, R. Sousa, and I. Domingues, "Ordinal data classification using kernel discriminant analysis: A comparison of three approaches," in *Proc. Int. Conf. Mach. Learn. Appl.*, 2012, pp. 473–477.

[15] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[16] W.-J. Li and D. Y. Yeung, "Localized content-based image retrieval through evidence region identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1666–1673.

[17] W.-J. Li and D. Y. Yeung, "MILD: Multiple-instance learning via disambiguation," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 1, pp. 76–89, Jan. 2010.

[18] R. El-Yaniv, D. Pechyony, and E. Yom-Tov, "Better multiclass classification via a margin-optimized single binary problem," *Pattern Recognit. Lett.*, vol. 29, no. 14, pp. 1954–1959, 2008.