Ordinal-Based Metric Learning for Learning Using Privileged Information

Shereen Fouad and Peter Tiňo

***Abstract*— Learning Using privileged Information (LUPI), originally proposed in [1], is an advanced learning paradigm that aims to improve the supervised learning in the presence of additional (privileged) information, available during training, but not in the test phase. We present a novel metric learning methodology that is specially designed for incorporating privileged information in ordinal classification tasks, where there is a natural order on the set of classes. This is done by changing the global metric in the input space, based on distance relations revealed by the privileged information. The proposed model is formulated in the context of ordinal prototype based classification with metric adaptation. Unlike the existing nominal version of LUPI in prototype models [8], [9], in ordinal classifications the proposed LUPI model takes explicitly into account the class order information during the input space metric learning. Experiments demonstrate that incorporating privileged information via the proposed ordinal-based metric learning can improve the ordinal classification performance.**

I. Introduction

Learning Using privileged Information (LUPI) is a new learning paradigm originally proposed by Vapnik [1] in a Support Vector Machine (SVM) framework, namely SVM+. It aims to improve the supervised learning in the presence of additional (substantial) information $x^* \in X^*$ about training examples $x \in X$, where the privileged information will not be available at the test stage. In the SVM+ context, the additional information is used to estimate the slack variable model. Indeed, slack variables are needed only in the training stage, eliminating the need for privileged information during testing. However, this approach **a)** is limited to binary classification problems; **b)** is difficult to interpret in terms of how exactly the additional information influences the resulting classifier through the slack model; **c)** can be computationally expensive for large-scale data sets, and **d)** is specially designed for the SVM model and hence inapplicable to other classifiers.

A more direct and flexible methodology for LUPI, based on Information Theoretic Metric Learning (ITML) [10], was proposed in the context of prototype-based classification [8], [9]. Prototype-based models lend themselves naturally to multi-class problems, are more amenable to interpretations and can be constructed at a smaller computational (compared to SVM+). The main idea behind the LUPI approach in [8], [9] is the modification of the metric in the original data space X, via the ITML formulation, based in data proximity 'hints' obtained from the privileged information space X^*. Furthermore, in [8], [9] two methods were proposed for incorporation of the new metric (obtained based on privileged information) into the original data space X.

Shereen Fouad and Peter Tiňo are with in the School of Computer Science, The University of Birmingham, Birmingham B15 2TT, United Kingdom, (email: shereen_afouad@hotmail.com, P.Tino@cs.bham.ac.uk).

All previous LUPI variants (whether in SVM+ or in metric learning formulation) were designed for incorporating privileged information for nominal classification problems. However, the training examples may be labeled by classes with a natural order imposed on them (e.g classes can represent rank). In the context of LUPI in prototype-based models [8], [9], the applied metric learning (i.e. ITML) for LUPI learns a distance metric for data space X from a number of (dis)similarity constraints obtained in the privileged space X^* through proximity information and label agreement. The appropriate metric for space X is found by keeping similar and dis-similar pairs closer and farther, respectively. Such an intuitive strategy may not, however, work well when classes are ordered, i.e. ordinal classification tasks. The ordinal label information is not considered explicitly during the constraints selection and metric learning. This can have detrimental effect on model performance.

This paper proposes an ordinal version of the ITML approach, namely Ordinal-based ITML (OITML), specifically designed for incorporating privileged data *during training* in ordinal prototype-based models. In particular, the new metric learning method will be applied in the context of Ordinal Generalized Matrix LVQ (OGMLVQ) [2], [3]. The proposed metric learning method, OITML , aims to learn a new metric in the original data space X, based on distance relations revealed in the privileged space X^*, *while preserving the linear order of classes* in the training set. The class order information is utilized in formulating the (dis)similarity constraints, as well as in the distance metric learning Itself. The new metric is then incorporated into X in the context of OGMLVQ classification as suggested in [8], [9].

In supervised nominal classification settings, recent advances in metric learning make it possible to learn distance functions that help to improve the classification accuracy, provided some side information is available (e.g. in the form of (dis)similarity constraints), e.g. Information Theoretic Metric Learning (ITML) [10], Large Margin Nearest Neighbor [11]. However, in this paper we are interested in learning a distance metric in the original space using similarity constraints obtained form the privileged space, that is specifically tailored to ordinal classification settings.

Some advances have been made in the development of metric learning algorithms for improving rank prediction accuracy (e.g. [17], [18], [19], [20]). Using a SVM formulation, the method presented in [20] aims to learn a metric from relative comparisons. The learned metric preserves ranks of

distances based on a set of qualitative constraints derived from the training data. Such constraints lead to a convex quadratic programming problem. A similar approach has been presented in [18], in the context of image retrieval. It addresses the problem of heterogeneous input space where 'must-link' (or similarity) constraints may vary from one query to another.

We empirically study our general methodology - LUPI via the proposed OITML - in three experimental settings: **a)** ordinal classification benchmark data sets, **b)** large scale astronomical ordinal classification problem and **c)** large scale ordinal time series prediction.

This paper has the following organization: Section II gives the background and briefly describes previous methods related to this study. Section III and IV introduces a novel ordinal-based metric learning approach for incorporation of privileged knowledge in ordinal prototype-based classification. Experimental results are presented in section V. Section VI concludes the study by summarizing the key contributions.

II. Background and Related Work

A. Prototype-Based Models and their Ordinal Extension

Prototype-based models, and particularly Learning Vector Quantization (LVQ) frameworks, constitute a family of supervised multi-class learning algorithms that adapt class prototypes to the training data in an on-line manner [6], [7], [5], [4]. Kohonen introduced the original LVQ1 scheme back in 1986 [6]. The prototypes are updated using Hebbian online learning. Numerous modifications/extensions of the basic LVQ1 scheme have been proposed in the literature. Recent variations allow for an explicit cost function [7] from which the prototype updates are derived, or allow for incorporation of adaptive distance measure in the data space [5], [4].

Most of the existing LVQ variants focus on predicting data labels from nominal (non-ordered) classes. Pattern recognition problems of classifying examples into ordered classes (ordinal classification), have received a great deal of attention as they appear in many practical applications (e.g. information retrieval [15], astronomical analysis [14] and medical applications [16]). Recently, the Generalized Matrix LVQ (GMLVQ) [4] was extended to the Ordinal GMLVQ (OGMLVQ) [2], that is specifically designed for classifying data items into ordered classes. The GMLVQ algorithm is a new heuristic LVQ extension of the GLVQ [7] model with a full generalized matrix tensor-based distance measure in the data space.

In the OGMLVQ [2], we assume training data of the form $(x_i, y_i) \in \mathbb{R}^m \times \{1, ..., K\},\ i = 1, 2, ..., n$, where the K classes are ordered, where $K > K-1 > ... > 1$. As in LVQ models, the OGMLVQ network consists of a number of prototypes $w_i \in \mathbb{R}^m$, $i = 1, 2, 3, ..., L$, which are characterized by their location in input space and their class label $c(w_i) \in \{1, ..., K\}$. Given an $(m \times m)$ positive definite matrix $\mathbf{\Lambda}$, the algorithm uses a generalized form of the full matrix distance measure,

$$d^2_{\mathbf{\Lambda}}(x_i, w) = (x_i - w)^T \mathbf{\Lambda} (x_i - w). \tag{1}$$

Positive definiteness of $\mathbf{\Lambda}$ can be achieved by substituting $\mathbf{\Lambda} = \mathbf{\Omega}^T \mathbf{\Omega}$, where $\mathbf{\Omega} \in \mathbb{R}^{m \times m}$, is a full-rank matrix. Furthermore, $\mathbf{\Lambda}$ needs to be normalized after each learning step to prevent the algorithm from degeneration.

The classification is based on a winner-takes-all scheme: a pattern $x_i \in \mathbb{R}^m$ is classified with the label of the closest prototype, $c(x_i) = c(w_j), \quad j = \arg\min_l d^2_{\mathbf{\Lambda}}(x_i, w_l)$. For each training pattern x_i, the algorithm implements Hebbian updates for the closest prototype w and for the metric parameter $\mathbf{\Omega}$. The nominal GMLVQ algorithm aims to position the class prototypes in the input space so that the overall misclassification error is minimized. However, the OGMLVQ model adapts the class prototypes so that the average absolute error ($|c(x_i) - c(w_j)|$) of class mislabeling is minimized. If $c(x_i)$ and $c(w)$ are close enough in their class order relation, then w is attracted towards x_i, otherwise w is repelled away. Thus, unlike in nominal LVQ models, the class order information is utilized during training in selection of the class prototypes to be adapted, as well as in determining the exact manner in which the prototypes get updated.

Note that missing values in training patterns can be handled by LVQ models. One of the most straightforward options is to simply ignore the missing dimensions when comparing prototypes with input data. The prototype and metric updates only affect the known features.

B. Information Theoretic Approach for Learning Using Privileged Information

In Learning Using privileged Information (LUPI) framework introduced in [1] in the context of SVM+, during training, a classifier may be given, along with training input $x_i \in X$, some additional information $x^*_i \in X^*$ about x_i. Such additional (privileged) information, however, will not be available in the test phase, where labels must be estimated using the trained model for previously unseen inputs $x \in X$ only (without x^*).

Our earlier work in [8], [9] introduced a new variant of LUPI, based on Information Theoretic Metric Learning (ITML) [10], in prototype-based models (particularly in the GMLVQ [4]). Prototype-based models lend themselves naturally to multi-class problems, are more interpretable and can be constructed at a smaller computational cost when compared to SVM+. The utilized ITML model is a supervised metric learning model that aims to learn a distance metric by minimizing the relative entropy between two multivariate Gaussian distributions, where the first one encodes the distance metric to be learned (via Mahalanobis distance) and the second one a reference Gaussian distribution. The minimization is enforced subject to a set of (dis)similarity constraints obtained from the data set [10].

In the ITML for LUPI formulation [8], [9], we are given training data $(x_i, y_i) \in \mathbb{R}^m,\ i = 1, 2, ..., n$ in the original space X and additional information $x^*_i \in X^*$ for $r \leq n$

training examples $x_i \in X$, $i = 1, 2, ..., r$. A metric tensor $\boldsymbol{M}$ on space X defines the distance,

$$d^2_{\boldsymbol{M}}(x_i, x_j) = (x_i - x_j)^T \boldsymbol{M}(x_i - x_j), \quad x_i, x_j \in X. \quad (2)$$

The privileged information in X^* is used to describe a set of constraints specifying pairs of privileged examples that are 'similar' (S_+) or 'dis-similar' (S_-):

- S_+ = $\{(x_i, x_j)|x_i$ and x_j are judged to be similar$\}$
- S_- = $\{(x_i, x_j)|x_i$ and x_j are judged to be dis-similar$\}$

Such constraints are imposed on the original space X through the ITML model. The metric $d^2_{\boldsymbol{M}}$ is modified so that the distances under the new learned metric $d^2_{\boldsymbol{C}}$ on X are shrunk and enlarged for pairs of points that have 'similar' and 'dis-similar' privileged information, respectively.

Whereas in typical ITML [10] (dis)similarity constrains are taken directly from label agreements between training pairs, the ITML for LUPI [8], [9] uses proximity information about privileged pairs in space X^*, as well as the label matches of corresponding pairs in space X. Given a global metric tensor $\boldsymbol{M}^*$ on X^* with the corresponding distance,

$$d^2_{\boldsymbol{M}^*}(x^*_i, x^*_j) = (x^*_i - x^*_j)^T \boldsymbol{M}^*(x^*_i - x^*_j), \;\; x^*_i, x^*_j \in X^*, \; (3)$$

the sets S_+ and S_- are constructed as follows:

- If $d^2_{\boldsymbol{M}^*}(x^*_i, x^*_j) \leq u^*$ and $c(x_i) = c(x_j)$(same label), then $(x_i, x_j) \in S_+$.
- If $d^2_{\boldsymbol{M}^*}(x^*_i, x^*_j) \geq l^*$ and $c(x_i) \neq c(x_j)$(different label), then $(x_i, x_j) \in S_-$.

Where u^* and l^* are small and large distance thresholds defined on space X^*, initialized as given in [8], [9].

In ITML for LUPI, closeness relation between the learned metric tensor $\boldsymbol{C}$ and the original metric tensor $\boldsymbol{M}$ is measured through the Bregman divergence (Burg) which is minimized during learning while enforcing the derived constraints.

Two approaches for incorporating the learned metric tensor $\boldsymbol{C}$ into a classifier operating on X were proposed in [8], [9]. The first approach performs a linear projection in the original space X such that distances pairs in S_+ are minimized, while they are maximized for pairs in S_-. The classifier is then trained on the transformed points. In the second approach, that is specially designed for the OGMLVQ classification, the new metric tensor $\boldsymbol{C}$ is used only for retraining the prototype positions in X, knowing that the metric tensor on X has changed. This is done by running the OGMLVQ algorithm while fixing $\boldsymbol{C}$.

It was shown in [8], [9] that the new learned metric (which reflect the privileged data distance structure) can improve the performance of nominal classification tasks. However, the training examples may be labeled by classes with a natural order imposed on them. The next section proposes an ordinal metric learning algorithm, based on the ITML [10], specifically designed for incorporating privileged for ordinal classification tasks.

III. Ordinal-based Information Theoretic Metric Learning (OITML) for Incorporating privileged Information

Consider a training data set $(x_i, y_i) \in \mathbb{R}^m \times \{1, ..., K\}$, where $i = 1, 2, .., , n$, and K is the number of ordered classes $K > K - 1 > ... > 1$. As before, assume that additional (privileged) information $x^*_i \in X^*$ is given for $r \leq n$ training examples $x_i \in X$, $i = 1, 2, ..., r$.

As in the case of nominal version of ITML for LUPI [8], [9], the aim here is to learn a data metric $\boldsymbol{C}$ for the original space X informed by inter-point distances in the privileged X^* space. The privileged information in X^* is used to describe sets of similarity S_+ and dis-similarity S_- constraints, as defined in section II-B. However, due to the ordinal nature of the underlying training classes, the class order information will be explicitly taken into account in the constraints derivation, as well as in distance metric learning for the original space X.

A. (Dis)similarity Constraints Derivation

Consider a privileged pair $(x^*_i, x^*_j) \in X^*$ with distance $d^2_{\boldsymbol{M}^*}(x^*_i, x^*_j)$, see Eq.(3) and the corresponding original training pair $(x_i, x_j) \in X$ with distance $d^2_{\boldsymbol{M}}(x_i, x_j)$ Eq. (2). Whereas in nominal ITML for LUPI [8], [9] constrains are decided based on proximity information and label agreement, in the OITML instead of strict label agreement, we will use the absolute class difference,

$$H(x_i, x_j) = \mid c(x_i) - c(x_j) \mid \quad (4)$$

where $c(x)$ denotes the class label of x.

Given a "tolerable class difference threshold" $\kappa \geq 0$, defined on the range of the loss function[1], the (dis)similarity sets S_+ and S_- are now constructed as follows[2]:

- If $d^2_{\boldsymbol{M}^*}\left(x^*_i, x^*_j\right) \leq u^*$ and $H(x_i, x_j) \leq \kappa$ (close in their class order), then $(x_i, x_j) \in S_+$.
- If $d^2_{\boldsymbol{M}^*}\left(x^*_i, x^*_j\right) \geq l^*$ and $H(x_i, x_j) > \kappa$ (apart in their class order), then $(x_i, x_j) \in S_-$,

where, as before, u^* and l^* are 'small' and 'large' distance thresholds (on X^*), respectively.

B. Weighting Scheme for the Metric Learning

Unlike nominal ITML for LUPI, the OITML for LUPI aims to learn an optimal metric in space X where distances induced among similar/dis-similar data pairs preserve the natural order relation between their classes. Thus, the notion of similar/dis-similar data pairs vary according to the corresponding class differences. Loosely speaking, if the class of point x_1 is closer in class order to the class of x_2 than to the class of x_3, i.e. $H(x_1, x_2) < H(x_1, x_3) \leq \kappa$, then during the metric learning the 'force' pulling together x_1 and x_2 is larger the force applied to x_1 and x_3. Analogous principle applies the "repulsive force" applied on dis-similar pairs.

[1] in our case $[0, K - 1]$

[2] Note that it is not necessary for all training points in X to be involved pairs of points in S_+ or S_-.

In the following we will propose a weighting scheme[3] for the OITML for LUPI which controls the amount of distance updates imposed on data pairs. There are two distinct weighting schemes for similar and dis-similar points.

1) **Weighting two similar points in** $(x_i, x_j) \in S_+$:
We propose a Gaussian weighting scheme,

$$\vartheta_{ij}^{+} = \exp\left\{-\frac{(H(x_i,x_j))^2}{2\sigma_+^2}\right\}, \tag{5}$$

where, σ_+ is the Gaussian kernel width.

2) **Weighting two dis-similar points in** $(x_i, x_j) \in S_-$:
Denote by ϵ_{max} the maximum class rank difference within all dis-similar pairs $(x_l, x_m)\,\forall (l,m) \in S_-$, i.e.,

$$\epsilon_{max} = \max_{(x_l,x_m)\in S_-} H(x_l, x_m)$$

The weight factor ϑ_{ij}^{-} for two dis-similar points $(x_i, x_j) \in S_-$ is then calculated as follows:

$$\vartheta_{ij}^{-} = \exp\left\{-\frac{(\epsilon_{max} - H(x_i,x_j))^2}{2\sigma_-^2}\right\} \tag{6}$$

where σ_- is the Gaussian kernel width.

The calculated weighting factors $\vartheta^{\pm}$ are utilized in the new OITML scheme presented in the next section.

C. Ordinal-Based Metric Learning

We aim to learn a new positive definite matrix (metric tensor) $\boldsymbol{C}$ on X, yielding the distance

$$d_{\boldsymbol{C}}^2(x_i,x_j) = (x_i - x_j)^T \boldsymbol{C}(x_i - x_j), \quad x_i, x_j \in X,$$

that while incorporating dominant distance relations in the privileged space X^*, also respects the class order.

Distance metric updates for similar/dis-similar pairs in space X are performed using the corresponding weights $\vartheta^{\pm}$. Thus, different degree of attraction and repulsive forces (based on data pairs class order relations) are allocated among similar and dis-similar pairs, respectively.

As in the standard ITML [10], the similarity between two the metrics $\boldsymbol{C}$ and $\boldsymbol{M}$ is measured through the Bregman divergence (Burg) defined over the cone of positive definite matrices as,

$$D_{Burg}(\boldsymbol{C},\boldsymbol{M}) = tr(\boldsymbol{C}\boldsymbol{M})^{-1} - \log\det(\boldsymbol{C}\boldsymbol{M}) - m,$$

where tr denotes the trace operator and m is the data dimensionality. Hence, the learning task is posed as the following constrained minimization problem[4]:

$$\begin{aligned}
&\min_{\boldsymbol{C}\succ\mathbf{0}} D_{Burg}(\boldsymbol{C},\boldsymbol{M}), \quad \text{subject to} \\
&d_{\boldsymbol{C}}^2(x_i,x_j) \le l\cdot\vartheta_{ij}^{+}, \text{ if } (x_i,x_j)\in S_+, \qquad \text{and} \\
&d_{\boldsymbol{C}}^2(x_i,x_j) \ge u\cdot\vartheta_{ij}^{-}, \text{ if } (x_i,x_j)\in S_-,
\end{aligned} \tag{7}$$

where $0 < l < u$ are the small and large distance thresholds on X, respectively.

[3]A similar technique was originally introduced in [2] for ordinal prototype based models.

[4] We use the notation $\boldsymbol{C} \succ \mathbf{0}$ to signify that $\boldsymbol{C}$ is positive definite matrix

As in the original ITML formulation [10], [8], [9], in the OITML, for guaranteeing a feasible solution for $\boldsymbol{C}$, a trade-off parameter $\nu > 0$ is introduced governing the influence of the constraints (and hence the influence of the privileged information). Let $s(i,j)$ denote the index of the (i,j)-th constraint, and let ξ be a vector of slack variables, initialized to ξ_0, with components equal to l for similarity constraints and u for dissimilarity constraints. Then the optimization problem can be reformulated as:

$$\begin{aligned}
&\min_{\boldsymbol{C}\succ\mathbf{0},\xi} D_{Burg}(\boldsymbol{C},\boldsymbol{M}) + \nu\cdot D_{Burg}(diag(\xi), diag(\xi_0)) \\
&\text{subject to} \quad d_{\boldsymbol{C}}(x_i,x_j) \le \xi_{s(i,j)}\cdot\vartheta_{ij}^{+}, \text{ if } (x_i,x_j)\in S_+, \\
&\text{and} \qquad d_{\boldsymbol{C}}(x_i,x_j) \ge \xi_{s(i,j)}\cdot\vartheta_{ij}^{-}, \text{ if } (x_i,x_j)\in S_-.
\end{aligned} \tag{8}$$

Similarly to the original ITML model [10], [8], [9], optimizing (8) involves repeatedly projecting (Bregman projections) the current solution onto a single constraint, via the update:

$$\boldsymbol{C_{t+1}} = \boldsymbol{C}_t + \beta_t \boldsymbol{C}_t (x_{i_t} - x_{j_t})(x_{i_t} - x_{j_t})^T \boldsymbol{C}_t, \tag{9}$$

where x_{i_t} and x_{j_t} are data points associated with one of the (dis)similarity constraints from $S_\pm$ at time t and the learning rate β_t decreases in time [10], [8], [9]. The algorithm initializes $\boldsymbol{C}$ to the precision matrix of the data in X (Mahalanobis distance).

The OITML algorithm for LUPI can be summarized as follows:

- **Inputs and Initialization:**
  $\boldsymbol{X}$: original training data $n \times m$ matrix, where m is the dimensionality of the original data.
  $\boldsymbol{X^*}$: privileged training data $r \times q$ matrix, where q is the dimensionality of the privileged data.
  κ: tolerable class difference threshold, used in Eq. (4).
  σ^+, σ^-: Gaussian kernel widths, used in Eq. (5) and (6), respectively.
  ν: trade-off parameter, used in Eq. (8).
  Lower and upper distance threshold (l^*, u^*) and (l, u) in spaces X^* and X, respectively.
  1) Based on Eq.(3) and (4), derive sets of similar S_+ and dis-similar S_- constraints, as given in section III-A.
  2) $\forall\ (i,j) \in S_+$ or $(i,j) \in S_-$ do:
     – if $(x_i, x_j) \in S_+$, calculate the corresponding ϑ_{ij}^{+} based on Eq. (5).
     – if $(x_i, x_j) \in S_-$, calculate the corresponding ϑ_{ij}^{-} based on Eq. (6).
  3) Initialize $\boldsymbol{C}$ to the precision matrix of the data in X.
  4) $\xi_{s(i,j)} \leftarrow l$ for $(i,j) \in S_+$, $\xi_{s(i,j)} \leftarrow u$ for $(i,j) \in S_-$
  5) Solve optimization problem in Eq. (8) with repeatedly projecting (Bregman projections) the current solution onto a single constraint, via the update in Eq. (9).
- **Output:** New metric tensor $\boldsymbol{C}$ in X incorporating the privileged data.

IV. Incorporating privileged Information into the OGMLVQ

As in [8], [9], we suggest two approaches for incorporating the learned metric tensor $\boldsymbol{C}$ into the OGMLVQ classifier operating on X.

1. Linear Transformation on X (OITML-LT): Knowing that metric tensor $\boldsymbol{C}$ is found in the parametrized form $\boldsymbol{C} = \boldsymbol{U}^T\boldsymbol{U}$, then for any training point $x \in X$, $\tilde{x} = \boldsymbol{U}x$ is the image of x under the basis transformation $\boldsymbol{U}$. Distances imposed on similar or dis-similar data pairs will now in general be shrunk or expanded according to (dis)similarity constraints. The standard OGMLVQ algorithm is now applied to the transformed data $\{(\tilde{x}_1, y_1), ..., (\tilde{x}_n, y_n)\}$. Note that,the linear transformation approach allows for application of any suitable ordinal regression classifier.

2. Extended OGMLVQ (OITML-Ext): OGMLVQ is first run on the original training set without privileged information, yielding a global metric $d^2_{\boldsymbol{M}}$ (given by metric tensor $\boldsymbol{M}$) and a set of prototypes $w_j \in \mathbb{R}^m$, $j = 1, 2, ..., L$. Then, the OITML technique finds metric $d^2_{\boldsymbol{C}}$ on X, based on the privileged information, that will replace $d^2_{\boldsymbol{M}}$. Finally, OGMLVQ is run once more while fixing the metric tensor $\boldsymbol{C}$ and modifying the prototype positions.

V. Experiments and Evaluations

We perform experiments in three ordinal classification settings; **a)** ordinal classification benchmark data sets, **b)** large scale astronomical ordinal classification problem and **c)** large scale ordinal time series prediction. In each experiment, we evaluate the effectiveness of incorporating the privileged information, via the proposed OITML, against the state of art OGMLVQ (trained without privileged information) used as a baseline. Furthermore, to show flexibility the proposed OITML model, we also employ the SVM Ordinal Regression with IMplicit Constraints (SVOR-IMC) classifier [15] operating in the modified metric found by OITML. For computational feasibility only data from the first experiment was used.

Three evaluation metrics are utilized to measure accuracy of predicted class $\hat{y}$ with respect to true class y on a test set:

1) **Mean Zero-one Error (MZE)** - (misclassification rate) is the rate of incorrect classified patterns, $MZE = \frac{\sum_{i=1}^{v} I(y_i \neq \hat{y}_i)}{v}$. where v is the number of test examples and $I(y_i \neq \hat{y}_i)$ denotes the indicator function returning 1 if the predicate holds and 0 otherwise.
2) **Mean Absolute Error (MAE)** - the average absolute deviation of the predicted ranks from the true ranks, $MAE = \frac{\sum_{i=1}^{v} |y_i - \hat{y}_i|}{v}$.
   However, MZE and MAE are not suitable for ordinal classification problems with imbalanced classes. Macro-averaged Mean Absolute Error (MMAE) [21], is an evaluation measure which estimates the mean of the class-conditional mean performances of the classifier. It is more appropriate for evaluating a classifier performance under imbalanced classes.
3) **Macro-averaged Mean Absolute Error (MMAE)** [21] - macro-averaged version of Mean Absolute Error - it is a weighted sum of the classification errors across classes, $MMAE = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{y_i=k} |y_i - \hat{y}_i|}{v_k}$. where K is the number of classes and v_k is the number of test points whose true class is k.

In all experiments, metric tensor $\boldsymbol{M}^*$ in X^* is set to the precision matrix[5] of the privileged training points $x^*_1, x^*_2, ..., x^*_r$ (Mahalanobis distance in X^*). The same applies to the initial metric tensor $\boldsymbol{M}$ in the original space X.

For OITML[6] approach, in order to set small and larger distance thresholds ($0 < l^* < u^*$) in space X^*, we first calculate all pairwise squared distances $d_{\boldsymbol{M}^*}(x^*_i, x^*_j)$, $1 \leq i < j \leq r$. These distances are then sorted in ascending order and, given a lower percentile parameter $a^* > 0$, a distance threshold l^* is found such that a^* percent of the lowest pairwise squared distances $d_{\boldsymbol{M}^*}(x^*_i, x^*_j)$ are smaller than l^*. Analogously, given an upper percentile parameter $b^* > a^*$, a distance threshold $u^* > l^*$ is found such that $(1 - b^*)$ percent of the largest pairwise squared distances $d_{\boldsymbol{M}^*}(x^*_i, x^*_j)$ are greater than u^*. The same strategy is applied for obtaining the distance thresholds $0 < l < u$ on X. In all experiments, (hyper-)parameters of OITML and OGMLVQ algorithms were tuned via cross-validation on the training set. In OITML, lower and upper bound percentiles for the privileged and original spaces are chosen over the values of $\{3, 5, 10\}$ for (a, a^*) and of $\{90, 95, 98\}$ for (b, b^*). Furthermore, the trade-off parameter ν is tuned over the values $\{0.01, 0.1, 1\}$ and the tolerable class difference threshold κ is is tuned over the values $\{0, 1, 2\}$.

For the OGMLVQ classifier, number of prototypes per class are tuned over the set $\{1, 2, 3, 4, 5\}$ in the first experiment (small-scale benchmark data sets), and over the set $\{5, 10, 15, 20\}$ in second and third experiments (large-scale data sets). The class prototypes are initialized as means of random subsets of training samples selected from the corresponding class. Relevance matrices are normalized after each training step to $\sum_i \boldsymbol{\Lambda}_{ii} = 1$ (see section II-A).

A. Controlled Experiments on Benchmark Data Sets

In this section we report on experiments performed using two benchmark ordinal regression data sets[7], namely Pyrimidines and MachineCpu, used in several ordinal regression formulations (e.g. [15]). Each data set was randomly partitioned into training/test splits 10 times independently, yielding 10 re-sampled training/test sets of size 50/24 and 150/59 for Pyrimidines and MachineCpu, respectively. On each data set, labels are discretized into five ordinal quantities using equal-frequency binning.

In order to demonstrate the advantage of the proposed method for incorporating the privileged information, an ini-

[5]The inverse of the covariance matrix.

[6]We modified the ITML Matlab code available from `http://www.cs.utexas.edu/users/pjain/itml/`. The parameters are tuned via cross-validation.

[7]Available at `http://www.gatsby.ucl.ac.uk/~chuwei/ordinalregression.html`

tial experiment is conducted which categorizes the input dimensions into 'original' and 'privileged' features in spaces X and X^*, respectively. Features categorization is driven by a 'wrapper' approach. For each data set, we sort the input features in terms of their relevance for the ordinal classifier (in our case OGMLVQ). The first most relevant half of the features will form privileged information, the remaining half will constitute the original space X. privileged features will only be incorporated in the metric learning, via the proposed OITML model, and will be absent during the ordinal classification testing. On each data set, parameters of the algorithm were tuned through 5-fold cross-validation on the training set. Note that, the ordinal classification process on the new manipulated metric in the input space does not necessarily need to be implemented using the OGMLVQ classifier. Hence, in this experiment (using the same data preprocessing and experimental settings), the proposed OITML is assessed using the SVOR-IMC classifier after learning the new metric via OITML. We used 5-fold cross validation to determine the optimal values of the SVOR-IMC model parameters (the Gaussian kernel parameter and the regularization factor) [15], both ranging from $\{$-2, -1,..., 1, 2$\}$.

The average MZE and MAE results over 10 randomized data sets splits (trials), along with standard deviations are shown in Table I. In the OGMLVQ classification, the OITML-LT approach achieves the best overall performance. In relative terms, on average, it outperforms the baseline OGMLVQ (trained on X only) by 8% and 6% on Pyrimidines and MachineCpu data sets, respectively. For the SVOR-IMC classification, incorporating the privileged information via the proposed OITML-LT improves the general performance on the Pyrimidines data set by 2% (relatively) when compared to the baseline SVMOR-IMC (trained on X only). However, it slightly reduces the performance on the MachineCpu data set. Incorporating privileged information via OITML in the OGMLVQ classifier is more successful than in the SVMOR based classifier. This is because OGMLVQ does not only incorporate the privileged information in terms of the learned metric on X, but it also re-positions the class prototypes 'optimally' with respect to the modified metric. Our OITML method can be considered a natural extension of the recent developments in LVQ, where the original LVQ approaches have been first extended to diagonal [5] and later to full metric tensors [4], which is further extended to the ordinal version, the OGMLVQ classifier [2].

TABLE I

MZE AND MAE RESULTS ON TWO BENCHMARK ORDINAL REGRESSION DATA SETS, ALONG WITH STANDARD DEVIATIONS ($\pm$) ACROSS 10 TRAINING/TEST RE-SAMPLING, FOR OGMLVQ AND SVOR-IMC CLASSIFICATIONS (WITHOUT PRIVILEGED DATA) AND THE OGMLVQ AND SVOR-IMC (WITH OITML FOR LUP).

Algorithm	Pyrimidines		MachineCpu	
	MZE	MAE	MZE	MAE
OGMLVQ	0.594 ±(0.063)	0.787 ±(0.082)	0.463 ±(0.059)	0.518 ±(0.066)
OITML-LT + OGMLVQ	0.548 ±(0.052)	0.728 ±(0.088)	0.429 ±(0.040)	0.496 ±(0.048)
OITML-Ext + OGMLVQ	0.587 ±(0.044)	0.749 ±(0.075)	0.424 ±(0.040)	0.501 ±(0.057)
SVOR-IMC	0.534 ±(0.056)	0.681 ±(0.12)	0.523 ±(0.026)	0.571 ±(0.038)
OITML-LT + SVOR-IMC	0.514 ±(0.101)	0.671 ±(0.18)	0.535 ±(0.019)	0.581 ±(0.044)

B. Galaxy Morphological Ordinal Classification using Full Spectra as privileged Information

Astronomers have been using several schemes for classifying Galaxies according to their morphological structure, i.e. visual appearance (e.g [12], [14]). The popular Hubble sequence scheme[8] classifies galaxies into three broad categories - *Elliptical, Spiral and Irregular*. Later on, the de Vaucouleurs scheme[9] (used in [14]) proposed a wider range of morphological classes (by considering more detailed morphological characteristic) which reflect galaxy age, thus imposing a meaningful order among the classes. This turns the galaxy morphology classification into an ordinal classification problem. Each class in the de Vaucouleurs system corresponds to one numerical value where smaller numbers correspond to early-type galaxies (e.g. elliptical and lenticular) and larger number correspond to late types (e.g. spiral and irregular).

Most of the existing galaxy morphological ordinal classification approaches use as input features galaxy photometric data, and ignore the costly-to-obtain full spectroscopic information. In a nominal classification setting (under the Hubble sequence classification scheme), our recent work on ITML for LUPI in prototype models [8], [9] revealed that using spectroscopic information as privileged information in the model construction phase (during training), alongside the original photometric data, can enhance the galaxy morphology classification based on photometric data only (test phase). This leads us to hypothesize that in the ordinal classification setting (under the de Vaucouleurs classification scheme), incorporating the spectral privileged information will improve the ordinal classification in test regime (using photometric data only).

Our data set contained 7,000 galaxies, classified into six ordinal morphological classes, extracted from a visual morphological classification catalog[10] in the Sloan Digital Sky Survey (SDSS) Data Release 4 (DR4) (galaxy IDs and their ordinal labels). As in [9], galaxies are represented through 13 photometric features (in X) and 8 privileged spectral features (in X^*), both extracted based on galaxy IDs from the SDSS DR9 [13] data catalog[11]. Algorithm parameters have been tuned through 10-fold cross-validation on the training set.

The MZE and MAE results, along with standard deviations

[8] http://www.galaxyzoo.org/

[9] http://en.wikipedia.org/wiki/Galaxy_morphological_classification

[10] http://vizier.cfa.harvard.edu/viz-bin/Cat?J/ApJS/186/427

[11] http://www.sdss3.org/dr9/

(10-fold cross validation) are shown in Table II. Note that the galaxy classes are almost balanced. As expected, in general, the inclusion of the spectral privileged information in the training phase via the OITML model enhances the ordinal classification performance, even though in the test phase the models are fed with the original photometric features only.

TABLE II

MZE AND MAE RESULTS ON THE ASTRONOMICAL DATA SET, ALONG WITH STANDARD DEVIATIONS ($\pm$) ACROSS 10 CROSS VALIDATION RUNS, FOR THE OGMLVQ (WITHOUT PRIVILEGED DATA) AND THE OGMLVQ (WITH OITML FOR LUPI)

Algorithm	MZE	MAE
OGMLVQ	0.458 ±(0.012)	0.648 ±(0.018)
OITML-LT + OGMLVQ	0.457 ±(0.018)	0.640 ±(0.019)
OITML-Ext + OGMLVQ	0.451 ±(0.018)	0.627 ±(0.012)

C. The Santa Fe Laser Time series Ordinal Prediction

In this experiment, we investigate the effectiveness of incorporating the privileged information (given in the form of future time series observations), via the proposed OITML, in an ordinal time series prediction problem. Our model is verified on the Santa Fe Chaotic Laser time series.

The Santa Fe Laser data set, obtained from a far-infrared-laser, is a cross-cut through periodic to chaotic intensity pulses of a real laser. The full time series[12], shown in Figure 1, consists of 10092 points. The laser activity produces periods of oscillations with increasing amplitude, followed by sudden, difficult to predict, activity collapses. A substantial

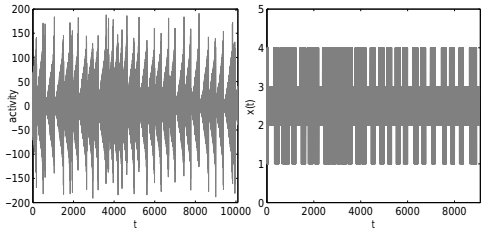

Fig. 1

THE SANTA FE LASER DATA SET (ON THE LEFT). TRANSFORMED SANTA FE LASER TIME SERIES (ORDINAL SYMBOLS)(ON THE RIGHT)

research activity has been devoted to the prediction and modeling of the Laser time series, e.g. [22]). However, this problem is studied here in the context of ordinal prediction settings rather than in nominal settings [25]. The model is predicting the order relations between the successive values instead of the time series values themselves, Ordinal prediction time series are found to be useful in several fields (e.g. analysis of stock prices and medical applications [24]). They are robust under non-linear distortion of the signal, since they use the ordinal relations of the time series rather than their real values.

As a pre-processing step, the laser activity changes have been quantized into ordinal symbolic streams. The method of extracting ordinal categorical information from complex time series forms the basis of ordinal symbolic dynamic [23]. The transformed time series is shown in Figure. 1

Given the chaotic laser time series y_t, $t = 1, 2, ..., 10092$, the differenced sequence $z_t = y_t - y_{t-1}$ has been quantized into a symbolic stream s_t, with s_t representing ordered categories of low/high positive/negative laser activity changes [25]:

$$s_t = \begin{cases} 1 & \text{(extreme down)} \quad \text{if } z_t \leq \Theta_1 \\ 2 & \text{(normal down)} \quad \text{if } \Theta_1 < z_t < 0 \\ 3 & \text{(normal up)} \quad \text{if } 0 \leq z_t < \Theta_2 \\ 4 & \text{(extreme up)} \quad \text{if } \Theta_2 \leq z_t, \end{cases} \tag{10}$$

where Θ_1 and Θ_2 correspond to Q percent (set here to 10%) and $(100 - Q)$ percent (set to 90%) sample quantile, respectively. Figure 2 plots the histogram of the differences between the successive laser activations. Dotted vertical lines show the corresponding cut values. Given the quantized laser

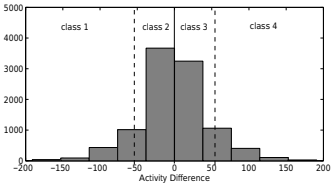

Fig. 2

HISTOGRAM OF THE DIFFERENCE BETWEEN THE SUCCESSIVE LASER ACTIVATION. DOTTED VERTICAL LINES SHOWS THE CUT VALUES $\Theta_1 = -56$ AND $\Theta_1 = 56$. ORDINAL SYMBOLS CORRESPONDING TO THE QUANTIZED REGIONS APPEAR ON THE TOP OF THE FIGURE.

time series, the task here is to predict the next laser activation change category $s(t+1)$, given the following (in the training):

- History of the last 10 activity differences $(z_{t-9}, z_{t-8}, ..., z_{t-1}, z_t)$, considered as the original training data in $X = \mathbb{R}^{10}$.
- 10 future activity differences $(z_{t+11}, z_{t+10}, ..., z_{t+2})$, considered as the privileged information in $X^* = \mathbb{R}^{10}$.

The first 5000 values of the series are used for training and validation, while the remaining 5092 are used for testing. Algorithm parameters have been tuned through 5-fold cross-validation on the training set.

The class distribution in the laser data set are highly imbalanced. Classes 2 and 3 (normal up/down) are more populated than classes 1 and 4 (extreme up/down). Therefore in Table III, along with MZE and MAE measures we also report the Macroaveraged Mean Absolute Error (MMAE), specially designed for evaluating classifiers operating on imbalanced data sets.

The results reveal that, in the context of ordinal time series prediction, the proposed formulation of incorporating

[12]Taken from `http://www-psych.stanford.edu/~andreas/Time-Series/SantaFe/A.cont`

TABLE III

MZE, MAE AND MMAE RESULTS ON THE SANTA FE LASER TEST SET FOR THE OGMLVQ (WITHOUT PRIVILEGED DATA) AND THE OGMLVQ (WITH OITML FOR LUPI)

Algorithm	MZE	MAE	MMAE
OGMLVQ	0.081	0.087	0.062
OITML-LT + OGMLVQ	0.073	0.078	0.052
OITML-Ext + OGMLVQ	0.071	0.077	0.054

the future time series data as privileged information [1], can lead to significant performance boost in the test regime over the standard classifier OGMLVQ trained on the historical observations only. Note that the inputs in the test phase were the same for both OGMLVQ and OGMLVQ-LT/Ext.

VI. CONCLUSION

We have introduced a novel ordinal-based metric learning methodology, based on Information Theoretic Metric Learning (ITML)[10], for Learning Using privileged Information (LUPI) in ordinal classifications. The proposed framework can be naturally cast in ordinal prototype based classification with metric adaptation (OGMLVQ) [2]. The privileged information is incorporated into the model operating on the original space X by changing the global metric in X, based on proximity relations obtained by the privileged information in X^*. We used two scenarios for incorporating the new learned metric on X in the ordinal prototype based modeling.

Unlike the nominal version of ITML for LUPI in prototype models [8], [9], in the proposed ordinal version the order information among the training classes is utilized to select the appropriate (dis)similarity constraints. Furthermore, the ordinal version of ITML realizes distance metric updates, for similar/dissimilar points in space X, using the assigned weights $\vartheta^{\pm}$, assigning different degree of similarity/dissimilarity measures (based on class order relations).

To our knowledge, this is the first work which studies the idea of LUPI into the ordinal classification setting.

We verified our framework in three experimental settings: **(1)** controlled experiments using two benchmark ordinal regression data sets, **(2)** a real world astronomical application-galaxy morphological ordinal classification. Here, the privileged information takes the form of costly-to-obtain full galaxy spectra. **(3)** ordinal time series prediction on chaotic time series. Experiment results revealed that incorporating privileged information via the proposed ordinal-based metric learning framework can improve the ordinal classification performance.

ACKNOWLEDGMENTS

Shereen's work was supported by the IDB. Peter's work was supported by a grant from the Biotechnology and Biological Sciences Research Council [BB/H012508/1].

REFERENCES

[1] V. Vapnik and A. Vashist: "A New Learning Paradigm: Learning Using privileged Information," *Neural Networks*, vol. 22, no. 5-6, pp.544–55, Elsevier Ltd, 2009.

[2] Sh. Fouad and P. Tino: "Adaptive Metric Learning Vector Quantization for Ordinal Classification," *Neural Computation*, vol. 24, no. 11, pp. 2825-2851, 2012.

[3] Sh. Fouad and P. Tino: "Prototype Based Modeling for Ordinal Classification," *Intelligent Data Engineering and Automated Learning*, vol. 7435, pp. 208-215, Lecture Notes in Computer Science, 2012.

[4] P. Schneider and M. Biehl and B. Hammer: "Adaptive Relevance Matrices in Learning Vector Quantization," *Neural Computation*, vol. 21, no. 12, pp. 3532–3561, 2009.

[5] B. Hammer and T. Villmann : "Generalized Relevance Learning Vector Quantization," *Neural Networks*, vol. 15, no. 8-9, pp. 1059–1068, 2002.

[6] T. Kohonen: "Learning Vector Quantization for Pattern Recognition," *Technical report*, No. (TKK-F-A601), Helsinki University of Technology. Espoo, Finland, 1986.

[7] A.S. Sato and K. Yamada: "Generalized Learning Vector Quantization," *Advances in Neural Information Processing Systems*, vol. 7, pp. 423–429, MIT Press, 1995.

[8] Sh. Fouad and P. Tino and S. Raychaudhury and P. Schneider: "Learning Using privileged Information in Prototype Based Models," *Artificial Neural Networks and Machine Learning*, vol. 7553, pp. 322-329, Springer, 2012.

[9] Sh. Fouad and P. Tino and S. Raychaudhury and P. Schneider: "Incorporating privileged Information Through Metric Learning," *IEEE Transactions on Neural Networks and Learning Systems*, Accepted for publication, February 2013.

[10] J. V. Davis and B. Kulis and P. Jain and S. Sra and I. S. Dhillon: "Information-Theoretic Metric Learning," *International Conference on Machine Learning*, pp. 209–216, ACM, 2007.

[11] K. Q. Weinberger and L. K. Saul: "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, JMLR.org, 2009.

[12] D. B. Wijesinghe and A. M. Hopkins and B. C. Kelly and N. Welikala and A. J. Connolly: "Morphological Classification of Galaxies and Its Relation to Physical Properties," *Monthly Notices of the Royal Astronomical Society*, vol. 404(4), pp. 2077-2086, 2010.

[13] C. P. Ahn and others: "The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey," *The Astrophysical Journal*, vol. 203(21), 2012.

[14] P. B. Nair and R. G. Abraham: "A Catalog of Detailed Visual Morphological Classifications for 14034 Galaxies in the Sloan Digital Sky Survey," *The Astrophysical Journal*, Supplement Series , vol. 186, pp. 427–456, 2010.

[15] W. Chu and S. S. Keerthi: "Support Vector Ordinal Regression," *Neural Computation*, vol. 19, no. 3, pp.792–815, 2007.

[16] J. S. Cardoso and J. F. Pinto da Costa and M. J. Cardoso: "Modelling Ordinal Relations with SVMs: An Application to Objective Aesthetic Evaluation of Breast Cancer Conservative Treatment," *Neural Networks*, vol. 18, no. 5-6, pp. 808–817, 2005.

[17] H. Ouyang and A. Gray: "Learning dissimilarities by ranking: from SDP to QP," *International Conference on Machine Learning*, pp. 728–735, ACM, 2008.

[18] J.E.Lee and R. Jin and A. K. Jain: "Rank-based distance metric learning: An application to image retrieval,"*Computer Vision and Pattern Recognition*, IEEE Conference, pp. 1–8, 2008.

[19] B. McFee and G. R. G. Lanckriet: "Metric Learning to Rank," *International Conference on Machine Learning*, pp. 775-782, 2010.

[20] M. Schultz, and T. Joachims: "Learning a Distance Metric from Relative Comparisons," *Neural Information Processing Systems*,vol. 16, pp. 41–48, MIT Press, 2004.

[21] S. Baccianella and A. Esuli and F. Sebastiani:"Evaluation Measures for Ordinal Regression," *Ninth International Conference on Intelligent Systems Design and Applications*, pp. 283–287, IEEE Computer Society, 2009.

[22] A.S. Weigend and N.A. Gershenfeld:"Time Series Prediction: Forecasting the Future and Understanding the Past," *Addison-Wesley*, 1993.

[23] K. Keller and M. Sinn:"Ordinal symbolic dynamics," *Technical Report*, no. A-05-14, 2005.

[24] G. Mller and C. Czado:"Regression Models for Ordinal Valued Time Series with Application to High Frequency Financial Data," 2002.

[25] P. Tino and G. Dorffner:"Predicting the Future of Discrete Sequences from Fractal Representations of the Past," *Machine Learning*, vol. 45, no. 2, pp. 187–217, Kluwer Academic Publishers, 2001.