Incorporating Privileged Information to Improve Manifold Ordinal Regression

M. Pérez-Ortiz, P. A. Gutiérrez and C. Hervás-Martínez
University of Córdoba, Dept. of Computer Science and Numerical Analysis, Rabanales Campus, Albert Einstein Building, 14071 Córdoba, Spain

Keywords: Manifold Learning, Ordinal Regression, Privileged Information, Kernel Learning.

Abstract: Manifold learning covers those learning algorithms where high-dimensional data is assumed to lie on a low-dimensional manifold (usually nonlinear). Specific classification algorithms are able to preserve this manifold structure. On the other hand, ordinal regression covers those learning problems where the objective is to classify patterns into labels from a set of ordered categories. There have been very few works combining both ordinal regression and manifold learning. Additionally, privileged information refers to some special features which are available during classifier training, but not in the test phase. This paper contributes a new algorithm for combining ordinal regression and manifold learning, based on the idea of constructing a neighbourhood graph and obtaining the shortest path between all pairs of patterns. Moreover, we propose to exploit privileged information during graph construction, in order to obtain a better representation of the underlying manifold. The approach is tested with one synthetic experiment and 5 real ordinal datasets, showing a promising potential.

1 INTRODUCTION

Ordinal regression is a learning task where the objective is to classify patterns into a set of predefined labels, but the labels include an order (Cardoso and da Costa, 2007; Chu and Keerthi, 2007; Li and Lin, 2007). For example, for an age estimation problem, people images could be classified into the classes $\{newborn, baby, young, adult, senior\}$. These categories are reflecting intervals of an actual latent variable (the real age of the person) but, contrary to standard regression, the latent variable is unobservable. On the other hand, the order between the categories makes this problem different from standard classification, and specific ordinal regression algorithms try to improve the quality of the classifier by introducing the order in the model and/or penalising the different classification errors (the magnitude of the error should be higher when the predicted class if further to the actual class) (Lin and Li, 2012).

Different methods have been proposed to deal with ordinal regression problems. Threshold models are one of the most popular approaches (McCullagh, 1980; Verwaeren et al., 2012), where the ordinal regression problem is formulated as the problem of estimating a real valued function and a set of $Q-1$ thresholds (Q is the number of classes), in such a way that one interval is assigned to each class ($[-\infty, b_1), [b_1, b_2), \ldots, [b_{Q-1}, \infty)$). This is the structure of the first specific model for ordinal regression, the proportional odds model (McCullagh, 1980), which is an ordinal version of binary logistic regression. Later on, nonlinear threshold models have appeared in the machine learning community, including different adaptions of other methods to the ordinal setting, such as support vector machines (R. Herbrich and Obermayer, 2000; Shashua and Levin, 2003; Chu and Keerthi, 2007), discriminant analysis (Sun et al., 2010) or Gaussian processes (Chu and Ghahramani, 2005). Other works decompose the original ordinal regression problem into several binary classification ones, by sequentially dividing the ordinal scale in binary labels (Frank and Hall, 2001; Cheng et al., 2008; Deng et al., 2010). Finally, a reduction framework can be found in (Cardoso and da Costa, 2007; Lin and Li, 2012), where ordinal regression is reduced to binary classification, but learning one single model for the binary problem where the input patterns are replicated, extended and weighted according to the ordinal label.

In this paper, we consider a manifold learning approach for ordinal regression. The idea of manifold

Pérez-Ortiz M., A. Gutiérrez P. and Hervás-Martínez C..
Incorporating Privileged Information to Improve Manifold Ordinal Regression.
DOI: 10.5220/0005075801870194
In *Proceedings of the International Conference on Neural Computation Theory and Applications* (NCTA-2014), pages 187-194
ISBN: 978-989-758-054-3

learning is to uncover the nonlinear structure embedded in a dataset, assuming that the high-dimensional observations lie on or close to an intrinsically low-dimensional manifold. There are different algorithms to learn this kind of structures, including the isometric feature mapping (Isomap) (Tenenbaum et al., 2000) or Laplacian eigenmaps (Belkin and Niyogi, 2001). Based on them, other manifold learning algorithms have been also proposed for classification, such as locality preserving projections (He and Niyogi, 2003) or the discriminant Laplacian embedding (DLE) (Wang et al., 2010).

In the context of ordinal regression, manifold learning has been considered in (Liu et al., 2011a; Liu et al., 2011b) based on the idea of preserving the intrinsic geometry of the data via the definition of a neighbourhood graph which also preserves the ordinal nature of the dataset. This graph is used to construct an adjacency matrix by using a generalised radial basis function. The Laplacian matrix is then derived and used for the learning process. A related method is proposed in (Liu et al., 2012), where several projections are iteratively computed. Finally, ranking on data manifolds is investigated in (Zhou et al., 2004), although the problem is defined as ranking, which is different from ordinal regression.

On the other hand, Vapnik and Vashist recently proposed a framework to apply support vector machines (SVM) to those cases where privileged information is available during the training phase, but not during test (Vapnik and Vashist, 2009). This kind of information can be found in many learning problems, where training samples present some special features which are not available during test because of their cost or simply because it is not possible. For example, suppose our goal is to find a rule that can predict outcome y of a treatment in a year given the current symptoms $\mathbf{x}$ of a patient. At the training stage, a doctor can also provide additional information $\mathbf{x}^*$ about the development of symptoms in three months, six months, and nine months (Vapnik and Vashist, 2009). The algorithm in (Vapnik and Vashist, 2009) was based on considering a slack model for this privileged information. Given that slacks are only considered during SVM optimisation and not included in the final model, their approach was able to benefit from this privileged information, mainly improving the convergence of the learning algorithm.

In this paper, we extend the ordinal regression manifold approach in (Liu et al., 2011b; Liu et al., 2011a) by considering privileged information during the neighbourhood graph construction. Under the assumption that privileged features are useful for the classification task, this approach would modify the neighbourhood structure to better represent the learning task. Moreover, we also consider a different approach for constructing the final distance matrix (by making use of the Dijkstra algorithm) and include this information into a kernel function, in order to apply support vector ordinal regression (Chu and Keerthi, 2007), as opposed to the ordinal discriminant-based projection method in the original proposal. Therefore, two main objectives can be found in this paper: Firstly, to analyse whether it is feasible to reformulate the notion of similarity for kernel functions when considering an ordinal manifold of the data and secondly, to study if the inclusion of privileged information helps to improve the constructed model. The approach is tested in one synthetic dataset and 5 real ones, showing a competitive performance.

The rest of the paper is organised as follows: Section 2 presents the methodology proposed, while Section 3 presents and discusses the experimental results. The last section summarises the main contributions of the paper.

2 METHODOLOGY

When dealing with multiclass classification, the goal is to assign an input vector $\mathbf{x}$ to one of Q discrete classes $\mathcal{C}_q, q \in \{1,\dots,Q\}$. To obtain the prediction rule $C : \mathcal{X} \to \mathcal{Y}$, we use an i.i.d. training sample $X = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ where N is the number of training patterns, $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $\mathcal{X} \subset \mathbb{R}^d$ is the d-dimensional input space and $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K\}$ is the label space. We are also provided a test set to obtain a reliable estimation of the classification error, $X_\mathrm{t} = \{\mathbf{x}_{\mathrm{t}i}, y_{\mathrm{t}i}\}_{i=1}^{N_\mathrm{t}}$, where N_t is the number of test patterns and $\mathbf{x}_{\mathrm{t}i} \in \mathcal{X}$, $y_{\mathrm{t}i} \in \mathcal{Y}$. Finally, many learning problems present some features which are available during training but not in the test phase. This privileged information complements training data in such a way that the training sample is $X = \{\mathbf{x}_i, \mathbf{x}_i^*, y_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$, $\mathbf{x}_i^* \in \mathcal{X}^*$, $y_i \in \mathcal{Y}$ and $\mathcal{X}^* \subset \mathbb{R}^{d^*}$ is the d^*-dimensional input privileged space. The test set is the same, given that privileged information is not available when applying the classifier.

Ordinal regression or ordinal classification are those problems where patterns have to be classified into naturally ordered labels. Consequently, the definition of this kind of problems is similar to the one introduced in the previous paragraph, but incorporating the following constraint: $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_K$, where $\prec$ denotes this order information.

Considering this ordering scale, one of the main hypothesis in ordinal regression is that the distance to adjacent classes is lower than the distance to non-

adjacent classes. Therefore, it can be said that ideally there exists a latent distance-based manifold of the output variable that results in $\mathcal{C}_q$ lying in the space between $\mathcal{C}_{q-1}$ and $\mathcal{C}_{q+1}$. In this paper, we test two different hypotheses. On the one hand and motivated by the large amount of ordinal kernel methods in the literature (Chu and Ghahramani, 2005; Chu and Keerthi, 2007; Sun et al., 2010; Liu et al., 2012), we test whether it is possible to include the manifold structure in the kernel matrix of kernel methods. Kernel matrices can be seen as structures of data that contain information about similarities among the patterns in a dataset. This notion of similarity is usually based on a distance relation between the patterns. Therefore, this distance can be modified to consider the manifold structure of the data. On the other hand, we test whether the inclusion of privileged information in the construction of the neighbourhood graph helps to improve the robustness and efficiency of the classification model. The following two subsections are related to the first hypothesis, while the last subsection covers the second one.

2.1 Constructing a Representative Graph for the Ordinal Manifold

This subsection comprises some elementary notions for constructing a representative graph for the ordinal manifold, which are used both in this paper and the previous work (Liu et al., 2011a; Liu et al., 2011b). Consider an undirected graph of N vertices, $G = (V,E)$, where V corresponds to the vertices of the graph and $E \subseteq [V]^2$ to the edges. In this case, the set of the training patterns form the set of vertices, $V = \{v_1, v_2, \ldots, v_N\} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, and the different edges connect pairs of patterns:

$$E = \{e_{i,j}\} = \{(v_i, v_j)\} = \{(\mathbf{x}_i, \mathbf{x}_j)\}, \qquad (1)$$

where $1 \leq i \leq N$ and $1 \leq j \leq N$. The set of edges is obtained via a k-neighbourhood analysis of the data, i.e. v_i is connected to v_j if $\mathbf{x}_i$ is one of the k-nearest neighbours of $\mathbf{x}_j$ *or* viceversa. Instead of this *or*, we could have considered the logical operator *and*, but we introduce this relaxed version of the neighbouring structure to prevent unconnected regions in the dataset. Note that if v_i is connected to v_j, there exist $e_{i,j}$ such that $e_{i,j} \in E$. For the purpose of constructing the neighbourhood graph, the Euclidean distance is used as the weight function (i.e. the one used for the neighbourhood analysis):

$$f(e_{i,j}) = d(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||_2, \qquad (2)$$

being $||\cdot||_2$ the L_2-norm operator.

As we aim to preserve the ordinal structure of the manifold, we could try to enlarge the locality between different ranks, as done in (Liu et al., 2011b). To do so, we can include a weight parameter w for the distances in such a way that these weights reflect the rank differences between data points:

$$w_{i,j} = |y_i - y_j| + 1. \qquad (3)$$

This weight information is applied to the distance function as follows: $d(\mathbf{x}_i, \mathbf{x}_j) = w_{i,j} \cdot ||\mathbf{x}_i - \mathbf{x}_j||_2$. The possibility of considering these weights is explored in the experiments of this paper (i.e. we consider both the weighted and unweighted versions of the proposal). Recall that this transformation of the distances is done before constructing the neighbourhood graph.

2.2 Including Graph Shortest Paths in the Kernel Matrix

Usually, for manifold learning algorithms, an adjacency matrix is used for the learning process (which is the underlying idea in (Liu et al., 2011a; Liu et al., 2011b)). In this paper, however, we try to analyse whether it is feasible to reformulate the notion of similarity for kernel functions when considering an ordinal manifold of the data. The main idea is to use the graph information obtained in the previous step to locate the different patterns in the underlying ordinal manifold of the data. To do so, we use the shortest path of the graph in order to provide a more smooth approach for the distances (as opposed to other manifold-based techniques where non-connected points are assumed to present an infinite distance).

In graph theory, the shortest path problem is the problem of finding a path between two vertices in a graph such that the sum of the weights of its constituent edges is minimised. As said, the constructed graph is undirected, so the notion of path is defined as a sequence of z vertices from v_1 to v_z, $p_{1z} = (v_1, v_2, \ldots, v_z) \in V^z$, such that v_i is adjacent to v_{i+1} for $1 \leq i < z$ (and therefore $e_{i,i+1}$ exists). Moreover, given a real-valued weight function $f : E \to \mathbb{R}$ (as said, the weighted or unweighted Euclidean distance) that assigns a cost to each edge and an undirected graph G, the shortest path from v to v' is the path $p_{1,z} = (v_1, \ldots, v_z)$ (where $v_1 = v$ and $v_z = v'$) that over all possible paths minimises the sum $\sum_{i=1}^{z-1} f(e_{i,i+1})$, where $e_{i,i+1} \in E$.

To compute the distance from one data pattern $\mathbf{x}_i$ to the rest but taking into account the manifold structure, we can compute the shortest paths from the vertex v_i to all the rest of vertices considering the well-known Dijkstra's algorithm (Dijkstra, 1959). Denote by P the set of paths obtained from this process, where

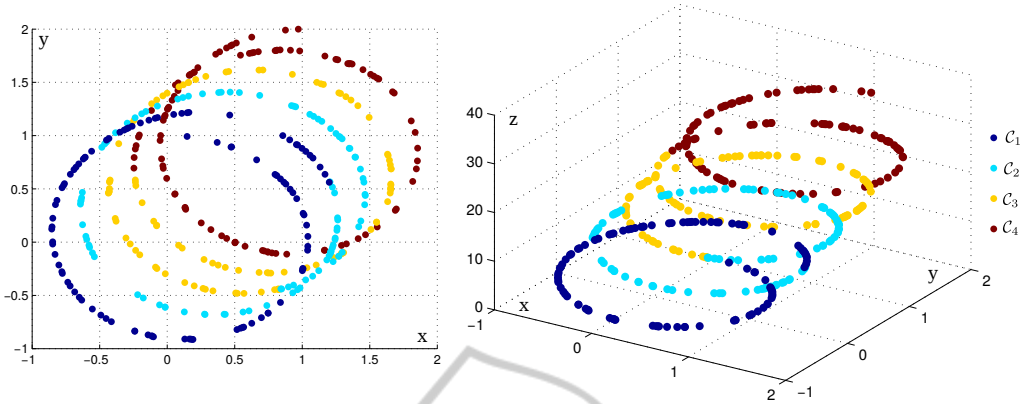

Figure 1: Representation of the `spiral` synthetic ordinal dataset. Left plot: Original dataset without privileged information. Right plot: Dataset including the privileged information as an additional feature. It can be seen that this privileged information improves the potential separability of the data.

$p_{i,j} \in P$ is the shortest path between v_i and v_j. Therefore, the distance from any two points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the training set is:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{h=1}^{z-1} f(e_{h,h+1}), v_1 = \mathbf{x}_i, v_z = \mathbf{x}_j. \quad (4)$$

where z is the length of the path between $\mathbf{x}_i$ and $\mathbf{x}_j$. Note that $d(\mathbf{x}_i, \mathbf{x}_j) = w_{i,j} \cdot ||\mathbf{x}_i - \mathbf{x}_j||_2$ if $\mathbf{x}_i$ is one of the nearest neighbours of $\mathbf{x}_j$. Therefore, to introduce the information of the location of each data point in the manifold in the kernel matrix, we modify the kernel function as follows:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2\sigma^2}\right), \quad (5)$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ is defined as in Eq. (4) and σ is the kernel parameter, as opposed to using the standard Gaussian kernel with the L_2-norm: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||_2^2}{2\sigma^2}\right)$. Note that this kernel matrix will still be positive semidefinite given that the only information changed is the distance function.

The kernel matrix obtained by this process is the one used for the training step. For the test phase, we first compute the distance from each test pattern $\mathbf{x}_{ti}$ to its nearest neighbour in training $\mathbf{x}_j$ and then, sum this distance to the shortest paths from $\mathbf{x}_j$ to the rest of training patterns. Consequently, the distance between a test point $\mathbf{x}_{ti}$ and all training points is:

$$d(\mathbf{x}_{ti}, \mathbf{x}_z) = d(\mathbf{x}_{ti}, \mathbf{x}_j) + d(\mathbf{x}_j, \mathbf{x}_z), \quad (6)$$

for $1 \leq i \leq N_t$ and $1 \leq z \leq N$.

This idea for the test phase corresponds to locate the test pattern in the graph and use the shortest paths information to compute the distance to the whole set of training patterns.

2.3 Including Privileged Information in the Graph

In order to motivate the inclusion of privileged information during manifold learning, Figure 1 represents a synthetic dataset presenting an ordinal manifold-based structure where the label of points is assigned according to the z coordinate. Data points lie on a leaning 3-dimensional spiral and labels are ordinal, with four classes $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ and $\mathcal{C}_4$. The Figure 1 also includes the projection over x and y coordinates. As can be seen, z coordinate is crucial to obtain a neighbourhood graph able to help in the ordinal classification task. Considering this z value as privileged information during graph construction would allow the classification of patterns, even when only x and y features are available during the test phase.

The privileged information can be easily included during distance calculation to construct a neighbourhood graph which takes into account this additional information. We can make use of the privileged features in the real-valued weight function f that assigns a value to edges of the graph:

$$\begin{aligned} f^*(e_{i,j}) &= ||(\mathbf{x}_i, \mathbf{x}_i^*) - (\mathbf{x}_j, \mathbf{x}_j^*)||_2 \quad (7) \\ &= \sqrt{\sum_{s=1}^{d} (x_{is} - x_{js})^2 + \sum_{s=1}^{d^*} (x_{is}^* - x_{js}^*)^2}. \end{aligned}$$

The whole process of neighbourhood analysis and shortest path computation is reformulated to work with this real-valued weight function. When considering this weight function, $f^*(e_{i,j})$, the distance function on Eq. (4) will be $d^*\left((\mathbf{x}_i, \mathbf{x}_i^*), (\mathbf{x}_j, \mathbf{x}_j^*)\right)$ and will be applied to the kernel function on Eq. (6). For the test phase, the privileged information is only consid-

ered for the graph that has been previously learnt, i.e.:

$$d\left(\mathbf{x}_{ti},(\mathbf{x}_z,\mathbf{x}_z^*)\right)=||\mathbf{x}_{ti}-\mathbf{x}_j||_2+d^*\left((\mathbf{x}_j,\mathbf{x}_j^*),(\mathbf{x}_z,\mathbf{x}_z^*)\right),$$

where $1\leq z\leq N$ and $\mathbf{x}_j$ is the closest training point from the test point evaluated $\mathbf{x}_{ti}$.

3 EXPERIMENTS

The proposed methodologies are based on generating a modified version of the kernel matrix (by exploiting the neighbourhood graph of the data), so they can be applied to any kernel classifier. In this way, we have considered the Support Vector Ordinal Regression with Implicit Constraints (SVORIM) (Chu and Keerthi, 2007), as it is one of the best performing threshold models for ordinal regression (Gutiérrez et al., 2012). 5 benchmark ordinal regression datasets have been used for the analysis, which are taken from publicly available repositories[1] (Asuncion and Newman, 2007; PASCAL, 2011). Additionally, a more controlled environment is provided by the `spiral` dataset, introduced in Section 2.3. Table 1 shows the characteristics of the evaluated datasets, where it can be checked that number of classes varies between 3 and 5.

In the experiments, we evaluate two different factors:

- The introduction of ordinal costs for penalising distances during the construction of the graph. Ordinal costs are based on the absolute cost. This factor will be used to confirm whether these costs are really useful for ordinal regression, as discussed in previous works (Liu et al., 2011b; Liu et al., 2011a).
- The improvement obtained by the privileged information. The graph will be constructed with and without privileged information to evaluate if the additional variables improve the quality of the model.

The most common evaluation measures for ordinal regression are the Mean absolute error (*MAE*) and the accuracy ratio (*Acc*) (Gutiérrez et al., 2012; Baccianella et al., 2009; Cruz-Ramírez et al., 2014). The *MAE* measure is used when the costs of different misclassification errors is not constant:

$$MAE=\frac{1}{N_{\mathrm{t}}}\sum_{i=1}^{N_{\mathrm{t}}}|y_{ti}-\hat{y}_{ti}|, \tag{8}$$

[1]Note that many of these datasets are frequently treated as nominal ones, without taking into account the order scale.

where $\hat{y}_{ti}$ is the label predicted for $\mathbf{x}_{ti}$. *MAE* values range from 0 to $Q-1$ (Baccianella et al., 2009).

Regarding the experimental setup, the datasets were divided 30 times using a holdout stratified technique with a 75% of the patterns for training and the remaining 25% for test. The splits of each holdout are the same for all the algorithms and one model is obtained for each training set and evaluated in the test set. The average test evaluation measures and the corresponding standard deviations are finally reported as the summary of the algorithm performance.

We use the standard Gaussian kernel for all the methods. Model selection is accomplished by cross-validating the hyperparameters of the algorithms considering only the training data (with a 5-fold cross-validation). The measure used to select the best parameter combination is *MAE*. The two parameters to be optimised are the kernel width (σ) and the cost parameter (C), both being selected within the values $\sigma, C \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. The number of nearest neighbours to be considered during graph construction is $k=3$. In those cases that a pattern is not connected to any other ones for the current value of k, we increase k until all patterns are connected to at least one.

For the `spiral` dataset, privileged information is the z coordinate. For the rest of datasets, we apply the Relief feature selection algorithm (Kira and Rendell, 1992) over the training set to sort the features by their relevance. We select half of the features (the most relevant ones) as privileged information ($\mathbf{x}^*$) and the rest as the original information ($\mathbf{x}$).

3.1 Results

Table 2 shows the test results for the 6 ordinal datasets considered in terms of *Acc* and *MAE*. The best result for each dataset is in bold face and the second one in italics. From this Table, we can outline several conclusions:

- When considering the ordinal weights, the *Acc* and *MAE* results are always improved by the privileged information. However, if the costs are not included, there are some datasets where the privileged information does not improve the results (`bondrate`, `contact-lenses` and `squash-unstored`). Given that the cross-validation criterion is the *MAE* (which is based on an absolute cost loss), we conclude that using these weights is necessary to properly obtain a benefit from the privileged information.
- From all the combinations, considering privileged information and ordinal weights is the best one,

Table 1: Characteristics of the six datasets used for the experiments: number of instances (Size), inputs (#In.), classes (#Out.) and patterns per-class (#PPC)

Dataset	Size	#In.	#Out.	#PPC
bondrate	57	37	5	(6,33,12,5,1)
contact-lenses	24	6	3	(15,5,4)
pasture	36	25	3	(12,12,12)
spiral	400	3	4	(50,50,50,50)
squash-unstored	52	52	3	(24,24,4)
tae	151	54	3	(49,50,52)

Table 2: Test results obtained for the different datasets (Mean± Standard Deviation of the 30 splits) by considering all the different manifold classification algorithms based on SVORIM.

		Acc		*MAE*	
		Privileged Information		Privileged Information	
Dataset	Ordinal Weights	No	Yes	No	Yes
bondrate	No	*57.28 ± 3.82*	56.54 ± 6.50	*0.6272 ± 0.0647*	0.6296 ± 0.0893
	Yes	56.54 ± 5.02	**58.52 ± 5.34**	0.6346 ± 0.0676	**0.6123 ± 0.0996**
contact-lenses	No	*61.11 ± 10.11*	*61.11 ± 10.11*	*0.5500 ± 0.0892*	*0.5500 ± 0.0892*
	Yes	58.89 ± 12.17	**62.22 ± 8.68**	0.5722 ± 0.1132	**0.5389 ± 0.0717**
pasture	No	*48.89 ± 14.76*	**51.85 ± 14.69**	*0.5370 ± 0.1600*	**0.5074 ± 0.1450**
	Yes	42.96 ± 15.91	43.70 ± 16.49	0.6037 ± 0.1668	0.6000 ± 0.1716
spiral	No	82.37 ± 4.16	*87.80 ± 2.70*	0.2260 ± 0.0589	*0.1867 ± 0.0505*
	Yes	85.03 ± 3.62	**87.90 ± 2.76**	0.2120 ± 0.0567	**0.1857 ± 0.0520**
squash-unstored	No	**52.56 ± 13.71**	50.77 ± 10.42	*0.4795 ± 0.1452*	0.4949 ± 0.1082
	Yes	49.74 ± 9.84	*51.54 ± 11.45*	0.5077 ± 0.0960	**0.4897 ± 0.1115**
tae	No	*35.35 ± 8.62*	**35.53 ± 8.40**	0.6570 ± 0.0867	*0.6526 ± 0.0783*
	Yes	34.91 ± 6.66	**35.53 ± 8.40**	0.6754 ± 0.0783	**0.6500 ± 0.0770**

obtaining the best results in four datasets and the second one in another.

- The most clear contribution of the privileged information is obtained for the spiral dataset. This is due to the fact in this more controlled environment data clearly belong to a low dimensional manifold and the class label is assigned according to the privileged information (z value). For the rest of datasets, the privileged information has been selected according to the Relief algorithm, which has known limitations. Nevertheless, there are some datasets where the contribution of privileged information is still quite noticeable (e.g. bondrate and contact-lenses).
- The original SVORIM algorithm (without using a manifold assumption) was run for the spiral dataset and the same configuration, leading to a performance of $Acc = 78.80 \pm 3.53$ and $MAE = 0.2617 \pm 0.0467$. It is noticeable that these values are worse than the ones obtained by the manifold proposals in this paper.

4 CONCLUSIONS

This paper considers a new approach to face ordinal regression problems based on manifold learning. This approach is based on constructing a neighbourhood graph with the purpose of obtaining the intrinsic structure of the data. The main paper contribution is that this neighbourhood graph can be improved by the use of privileged information, information that is available during training but not in the test phase.

The algorithm is applied to 5 ordinal classification real problems and one synthetic dataset. When combined with SVORIM, the results of this paper confirm that privileged information is able to improve generalisation results for almost all the cases considered. The distances used in the kernel matrices are obtained using the privileged features, which (under the assumption that privileged information is really informative) better reflects the data structure.

Several future research directions are still open from the work in this paper. First of all, more datasets should be considered, including datasets with a higher number of patterns and with a more clear manifold

structure. For example, the experiments considered in (Liu et al., 2011b) cover the UMIST face, MovieLens and the USPS datasets, which are known to contain an underlying manifold structure. The problem is that meaningful privileged information has to be found for these problems. Secondly, the methods should be compared against standard manifold classifiers to check their performance. Finally, alternative kernel methods apart from SVORIM could be considered together with the proposals in this paper.

ACKNOWLEDGEMENTS

This work has been subsidized by the TIN2011-22794 project of the Spanish Ministerial Commission of Science and Technology (MICYT), FEDER funds and the P11-TIC-7508 project of the "Junta de Andalucía" (Spain).

REFERENCES

Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

Baccianella, S., Esuli, A., and Sebastiani, F. (2009). Evaluation measures for ordinal regression. In *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications (ISDA 09)*, pages 283–287, Pisa, Italy.

Belkin, M. and Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591.

Cardoso, J. S. and da Costa, J. F. P. (2007). Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research*, 8:1393–1429.

Cheng, J., Wang, Z., and Pollastri, G. (2008). A neural network approach to ordinal regression. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN2008, IEEE World Congress on Computational Intelligence)*, pages 1279–1284. IEEE Press.

Chu, W. and Ghahramani, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041.

Chu, W. and Keerthi, S. S. (2007). Support vector ordinal regression. *Neural Computation*, 19(3):792–815.

Cruz-Ramírez, M., Hervás-Martínez, C., Sánchez-Monedero, J., and Gutiérrez, P. A. (2014). Metrics to guide a multi-objective evolutionary algorithm for ordinal classification. *Neurocomputing*, 135:21–31.

Deng, W.-Y., Zheng, Q.-H., Lian, S., Chen, L., and Wang, X. (2010). Ordinal extreme learning machine. *Neurocomputation*, 74(1-3):447–456.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

Frank, E. and Hall, M. (2001). A simple approach to ordinal classification. In *Proc. of the 12th Eur. Conf. on Machine Learning*, pages 145–156.

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

He, X. and Niyogi, P. (2003). Locality preserving projections. In *NIPS*, volume 16, pages 234–241.

Kira, K. and Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, pages 129–134.

Li, L. and Lin, H.-T. (2007). Ordinal Regression by Extended Binary Classification. In *Advances in Neural Inform. Processing Syst. 19*.

Lin, H.-T. and Li, L. (2012). Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 24(5):1329–1367.

Liu, Y., Liu, Y., and Chan, K. C. C. (2011a). Ordinal regression via manifold learning. In Burgard, W. and Roth, D., editors, *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, pages 398–403. AAAI Press.

Liu, Y., Liu, Y., Chan, K. C. C., and Zhang, J. (2012). Neighborhood preserving ordinal regression. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service (ICIMCS12)*, pages 119–122, New York, NY, USA. ACM.

Liu, Y., Liu, Y., Zhong, S., and Chan, K. C. (2011b). Semi-supervised manifold ordinal regression for image ranking. In *Proceedings of the 19th ACM international conference on Multimedia (ACM MM2011)*, pages 1393–1396, New York, NY, USA. ACM.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, 42(2):109–142.

PASCAL (2011). Pascal (pattern analysis, statistical modelling and computational learning) machine learning benchmarks repository.

R. Herbrich, T. G. and Obermayer, K. (2000). Large margin rank boundaries for ordinal regression. In Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*, pages 115–132. MIT Press.

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

Sun, B.-Y., Li, J., Wu, D. D., Zhang, X.-M., and Li, W.-B. (2010). Kernel discriminant learning for ordinal regression. *IEEE Transactions on Knowledge and Data Engineering*, 22:906–910.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Vapnik, V. and Vashist, A. (2009). A new learning

paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557.

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

Wang, H., Huang, H., and Ding, C. H. (2010). Discriminant laplacian embedding. In *AAAI*.

Zhou, D., Weston, J., Gretton, A., Bousquet, O., and Schölkopf, B. (2004). Ranking on data manifolds. In *Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems (NIPS2003)*, pages 169–176.