Knowledge-Based Systems 88 (2015) 144–153

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A novel ordinal learning strategy: Ordinal nearest-centroid projection

Qing Tian, Songcan Chen *

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 30 September 2014
Received in revised form 29 May 2015
Accepted 29 July 2015
Available online 1 August 2015

Keywords:
Ordinal regression
Ordinal nearest-centroid projection
Combinatorial optimization
Quadratic programming

ABSTRACT

Ordinal regression (OR) is a learning paradigm lying between classification and regression and has been attracting increasing attention in recent years due to its wide applications such as human age estimation. To date, there have been a variety of methods proposed for OR, among which the category of threshold-based OR becomes one of the representatives with preferable performance. Typical threshold-based methods, such as discriminant learning for OR (i.e., KDLOR), OR via manifold learning (i.e., MOR), usually seek an OR projection direction along which to maximally separate classes by a sequence of ordinal thresholds. Although having yielded encouraging results, they still leave a performance space that can be further improved since (1) the thresholds involved are optimized independently from each other, and (2) the ordinal constraints just associate with class means (or say centroids) which are generally under-represented for class distributions. Motivated by the analysis, in this work we propose to jointly learn the thresholds across samples and class centroids by seeking an optimal direction along which all the samples are distributed as in order as possible and maximally cater for nearest-centroid distributions, which we call *Ordinal Nearest-Centroid Projection* (OrNCP) and is formulated as a combinatorial optimization problem. For efficiency of optimization, we further relax the problem to a quadratic programming (QP-OrNCP) that in form covers the KDLOR and MOR as its special cases. Finally, through extensive experiments on synthetic and real ordinal datasets, we demonstrate the superiority of the proposed method, over state-of-the-art methods.

1. Introduction

In the community of machine learning, ordinal regression (OR) has been attracting increasing attentions due to its dual nature of *discrete regression* and *ordinal classification*, and especially wide applications in recommender system [1], web page ranking [13], image retrieval [32], medical image diagnosis [33], and facial age estimation [5,6].

To date, many varieties of methods have been proposed to implement OR, which mainly can be grouped into three categories. The main idea of the *first family* is to conduct OR straightforwardly by means of the off-the-shelf regressors. For example, Kramer et al. [15] first converted the ordinal labels into real values, and then borrowed the standard regressor for ordinal learning. However, an associated problem is that it is usually difficult to naturally measure the distance between the ordinal labels [4].

The methods of the *second category* reported in literature perform OR by referring to classification with single or multiple output codings. In [9], Frank et al. represented the ordinal labels with a batch of binary numbers, and then obtained the OR result by combining the outputs of many nested binary classifiers that are independent to each other. Along the line of Frank and Hall [9], Waegeman and Boullart [31] further assigned weights for the involved binary classifiers to promote the generalization ability. To better achieve the goal of ordinal learning with a single augmented binary classifier, Cardoso and De Costa [4] took the way of training data replication. And recently, Lin and Li [16] even unified the methods of this family through modeling with cost matrix.

In order to more naturally conduct OR learning in accordance with its nature of lying between the classification and the regression, the *third category* assumes that the naturally ordinal classes can be orderly separated by a sequence of monotonous thresholds along the projection direction, thus which is also called the *threshold-based*. Along this way, the proportional odds model (POM) proposed in McCullagh [20] is the first attempt by modeling a linear combination of the training data for ordinal output, which later was extended non-linearly by neural network [19] and kernel mapping [24], respectively. After the POM, Crammer et al. [8] introduced a set of separation thresholds to the perception algorithm to perform online OR learning. Shashua and Levin [27] carried out the OR learning based on the principle of fixed-margin

* Corresponding author.
E-mail addresses: tianqing@nuaa.edu.cn (Q. Tian), s.chen@nuaa.edu.cn (S. Chen).

http://dx.doi.org/10.1016/j.knosys.2015.07.037
0950-7051/

and sum-of-margin, respectively. Based on the work of Kramer et al. [15], Chu and Keerthi [7] developed ordinal variations of the SVM, called SVOR-IMC and SVOR-EXC, respectively associated with implicit and explicit ordinal thresholds. More recently, in order to generate a more discriminative OR estimator, Sun et al. [28] developed a discriminative OR learning model, KDLOR, by imposing order constraints between each two neighboring class means, and experimentally demonstrated the superiority of KDLOR to the SVORs. Considering the success of KDLOR, Pérez-Ortiz et al. [23] proposed to optimize the thresholds by means of maximum likelihood estimations; and in Tian et al. [30], Tian et al. incorporated the spatial information of images into the OR. Using the same form of ordinal constraints as in Sun et al. [28], Liu et al. [17] proposed the manifold ordinal regression (MOR) by preserving geometrical manifold embedded in data. More recently, Sun et al. [29] and Liu et al. [18], respectively, extended their methods to multidirectional versions, with the same form of ordinal constraints in each direction as in KDLOR or MOR. With a same goal of seeking for multidirectional OR projections, Gutiérrez et al. [12] even employed the complex neural networks to train a set of concentric hyperspheres for OR. Considering that in most cases unidirectional OR can be similarly extended to corresponding multidirectional counterpart, so in this work we just focus on the unidirectional OR. As typical unidirectional OR methods with preferable performance, although the KDLOR and MOR can yield more competitive results than other ones including the SVORs, they still mainly suffer from two problems:

1. Their ordinal constraints are built on a sequence of partial order constraints each of which just associates with one threshold between two neighboring classes. That is, the thresholds involved are optimized independently from each other such that the global optimality is difficult to guarantee.
2. The constraints involved associate with just class means (i.e., class centroids), which may be under-represented for the individual distributions of data classes, as demonstrated in Fig. 1.

Motivated by the above analysis, as well as the threshold-based OR decision rule (*i.e., actually, a test instance is assigned into the ordinal class whose centroid is nearest to the instance*), in this work we propose to jointly learn the ordinal thresholds across training samples and class centroids through seeking for an optimal projection direction along which all the samples are distributed as in order as possible and maximally cater for a nearest-centroid distribution for each class. As a result, such a learning problem can be formulated as a combinatorial optimization problem. For efficiency of implementation, we relax the problem to a quadratic programming that can be easily solved and makes the KDLOR and MOR become its reduced cases. Finally, through extensive comparative experiments on a toy problem, hand-written digit recognition and human facial age estimation, we demonstrate the superiority of our strategy in performing OR. To our knowledge, this is the first work to learning OR directly from the perspective of threshold-based OR decision rule.

Our main contributions in this paper are as follows:

- Propose a novel ordinal regression (OR) learning strategy, namely *Ordinal Nearest-Centroid Projection* (OrNCP), specially following the threshold-based OR decision rule and formulate it as a combinatorial optimization problem.
- For efficiency of implementation, relax the OrNCP to a quadratic programming problem (QP-OrNCP) that covers two typical OR approaches KDLOR and MOR as its special cases.
- Experimentally demonstrate the effectiveness and superiority of our strategy in performing OR, on both synthetic and real-world datasets, in which we also investigate the influences of the *form* and the *granularity* of the ordinal constraints on OR performance, finding that our ordinal nearest-centroid constraints imposed across class centroids and data samples are superior to those either just on class centroids or just on class samples.

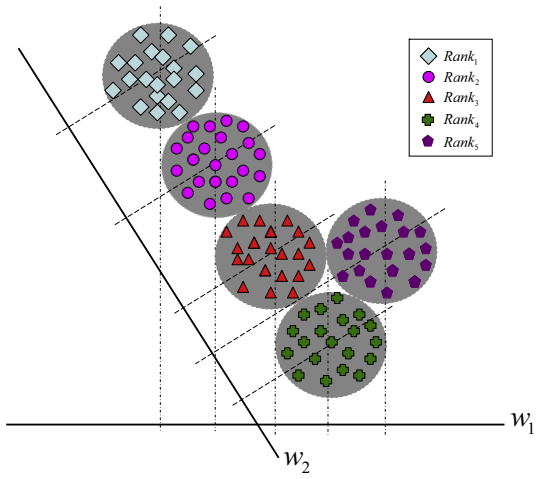


Fig. 1. Comparison between different OR projection directions, in which the direction of w_1 represents the one trained via MOR or KDLOR which encourages the OR learning to cater for their ordinal constraints built between class means of the data, whereas a more desirable direction may be the w_2 as it is relatively more consistent with the distribution trend of the classes.

The rest of this paper is organized as follows. In Section 2, we briefly review two typical threshold-based OR methods, i.e., the KDLOR and MOR, and especially show their strategy in preserving the order of data. In Section 3, we propose our learning strategy for OR, directly following the threshold-based OR decision rule. And to evaluate the effectiveness of our strategy, we conduct comparative experiments on synthetic and real world datasets in Section 4. Finally, we conclude the paper in Section 5.

2. Related work

Prior to introducing our work, we first give a brief review for the KDLOR and MOR, two typical threshold-based OR methods closely-related to our work, and analyze their limitations in OR learning.

2.1. KDLOR

In Sun et al. [28], an ordinal counterpart of discriminant learning (i.e., KDLOR[1]) was proposed for catering for ordinality of data classes. With the goal of minimizing the within-class scatter and simultaneously preserving the order structure among the data classes, Sun et al. designed the following formulation

$$\begin{aligned} &\min_{w,\rho} \quad w^T \cdot S_w \cdot w - C \cdot \rho \\ &s.t. \quad w^T \cdot (m_{k+1} - m_k) \geqslant \rho, \quad k = 1, 2, \ldots, K-1, \end{aligned} \tag{1}$$

where S_w denotes the whole within-class scatter matrix, C is a trade-off parameter, m_k represents the class mean of the kth class, and ρ denotes the margin between two neighboring classes after projected along the w.

[1] Unless otherwise specified, in this paper we do not explicitly distinguish between the linear and non-linear cases.

2.2. MOR

Based on the assumption that many observations usually comply with certain low-dimensional manifold distribution, Liu et al. [17] proposed to employ the manifold learning to OR, and presented the ordinal regression with manifold preservation, called MOR formulated as

$$\begin{aligned} &\min_{w,\rho} \quad w^T \cdot X \cdot L \cdot X^T \cdot w - C \cdot \rho \\ &s.t. \quad w^T \cdot (m_{k+1} - m_k) \geqslant \rho, \quad k = 1, 2, \ldots, K-1, \end{aligned} \tag{2}$$

in which X denotes the training data matrix, L is the Laplacian matrix as defined in Belkin and Niyogi [2], and the meanings of other symbols here are the same as those in Section 2.1.

2.3. Limitations of the KDLOR and MOR

Comparing Eqs. (1) and (2), it can be seen that the ordinal constraints of

$$w^T \cdot (m_{k+1} - m_k) \geqslant \rho, \quad k = 1, 2, \ldots, K-1, \tag{3}$$

are the same in the form, which restrict the projected means of classes to be arranged in order by combining a sequence of partial order constraints between the means of two neighboring classes. However, such type of constraints leaves a performance space that can be further improved, since that

1. The thresholds, lying between every two neighboring classes, just associate with their corresponding partial order constraint and are optimized independently from each other, by which the global ordinal optimality is difficult to guarantee.
2. The partial order constraints are built upon the means of classes, which are generally under-represented for distributions of data classes. As a result, the performance of OR learning will be deteriorated, as demonstrated in Fig. 1.

3. Proposed method

To mitigate the problems aforementioned, in what follows we propose a novel strategy to jointly learn the ordinal thresholds across training samples and class centroids (i.e., class means), motivated by the analysis of threshold-based OR decision rule.

3.1. Methodology: Ordinal Nearest-Centroid Projection (OrNCP)

Through the literature review about the threshold-based OR methods aforementioned in Section 1, we can summarize the main idea of threshold-based OR learning as follows:

- *In the training phase:* on the training set, seek for an optimal projection direction (here denoted as w) along which all the samples are arranged as consistent with their orders as possible and every two neighboring classes are separated by thresholds, which are constrained to be ordinal.
- *In the test phase:* for a new test instance, project it along the direction w, and compare its resulting projection with the OR thresholds to make an order decision.

From the analysis above, it can be seen that the threshold-based OR learning is quite similar to the nearest-centroid projection learning [10], as demonstrated in Fig. 2. More importantly, in the learning of the nearest-centroid projection, the distances of each sample to class-centroids of all classes are considered together. That is, in seeking for the optimal nearest-centroid projection direction, all the thresholds associated between classes are adjusted simultaneously.

Motivated by the analysis above, in what follows we propose a novel OR learning strategy to jointly learn the thresholds involved across samples and class centroids through seeking for an optimal direction along which all the samples are distributed as in order as possible and meanwhile maximally cater for a nearest-centroid distribution for each class, coined as OrNCP. More concretely, we assume that there are a set of $\{x_i\}_{i=1}^N$ samples associated with ordinal class labels $\{l\}_{l=1}^K$ from totally K classes and the kth class-set as $\{x_j^k\}_{j=1}^{N_k}$. In addition, let $\overline{X}_k$ denote the class mean of the kth class. Then, the optimization objective for the nearest-centroid projection, as demonstrated in Fig. 2, can be formulated as

$$\max_w \quad \frac{\sum_{k=1}^K \sum_{i=1}^{N_k} \sum_{p \neq k} \mathcal{I}[|w^T(x_i^k - \overline{X}_p)| > |w^T(x_i^k - \overline{X}_k)|]}{\sum_{k=1}^K N_k(K-1)}, \tag{4}$$

where $|\cdot|$ represents the *absolute value operator*, w denotes the optimal projection vector to seek and $\mathcal{I}[\cdot]$ is the indicator function which returns 1 if the argument is true and 0 otherwise. However, it can be seen that the ordinal relationship among the classes has not been reflected in Eq. (4) yet. Therefore, we reformulate (4) by taking the ordinal information into account to cater for the OR and thus propose our OR strategy, which we call ordinal nearest-centroid projection learning, or OrNCP for short, whose objective is formulated as

$$\begin{aligned} \max_w \quad & \frac{\sum_{k=1}^K \sum_{i=1}^{N_k} \sum_{l<k} w_{ikl} \cdot \mathcal{I}[w^T(x_i^k - \overline{X}_l) > k - l]}{\sum_{k=1}^K N_k(K-1)} \\ & + \frac{\sum_{k=1}^K \sum_{i=1}^{N_k} \sum_{h>k} w_{ikh} \cdot \mathcal{I}[w^T(\overline{X}_h - x_i^k) > h - k]}{\sum_{k=1}^K N_k(K-1)} \\ & + \frac{\sum_{k=1}^K \sum_{i=1}^{N_k} \mathcal{I}[|w^T(x_i^k - \overline{X}_k)| < 1]}{\sum_{k=1}^K N_k K}, \end{aligned} \tag{5}$$

where w_{ikl} (and similarly w_{ikh}) denotes the ordinal weight defined jointly on the ith sample from the kth class and the class centroids of the kth and pth classes. It can be found that by taking account of the ordinal relationships among data classes into learning, the nondirectional nearest-centroid projection learning of (4) is transformed into an ordinal counterpart of (5). More specifically, instead of selecting a projection for nondirectional nearest-centroid distribution, we propose to seek for an optimal projection along which all class samples are distributed as in order as possible and maximally cater for a nearest-centroid distribution for each class, as depicted by the first two terms and the third one of (5). It is worth to point out that it is these critical modifications that casts the ordinary nearest-centroid projection learning into an ordinal counterpart for OR, and by optimizing Eq. (5), the ordinal thresholds involved between the classes can be jointly learned, which, to our knowledge, is the first work to perform OR specifically following the threshold-based OR decision rule.

3.2. Relaxation of OrNCP: QP-OrNCP

Although the OrNCP formulated in Eq. (5) shows us a novel strategy to perform OR directly according to the threshold-based OR decision rule, the OrNCP itself is a bit difficult to be optimized since that it is a combinatorial optimization problem. In practice, it is extremely time-consuming to seek for the optimal solution, especially when the size of training set is large.

To reduce the computational complexity, we relax (5) by introducing the *hinge loss* function [11] and enforcing the with-class scatters to be compact to approximate the first two and the third indicator functions $\mathcal{I}[\cdot]$ in (5), respectively. As a result, we can reformulate (5) as

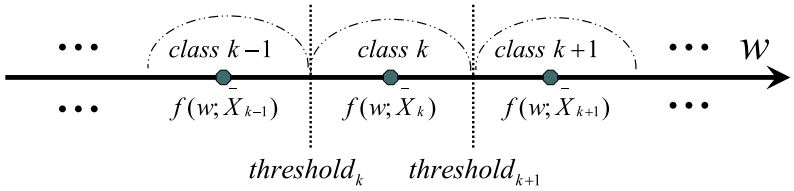


Fig. 2. Demonstration of nearest-centroid rule. Here, w denotes the projection vector, $f(w;\overline{X}_k)$ represents the new centroid of kth class after projection along w by function $f(\cdot)$ with $\overline{X}_k$ being the original class centroid, and $threshold_k$ is the separation threshold between the $(k-1)$th class and kth class.

$$\begin{aligned}\min_{w}\quad & \sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{l<k} w_{ikl}\cdot max\{(k-l)-w^T(x_i^k-\overline{X}_l),\ 0\}\\ & +\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{h>k} w_{ikh}\cdot max\{(h-k)-w^T(\overline{X}_h-x_i^k),\ 0\}\\ & +\sum_{k=1}^{K}\sum_{i=1}^{N_k}\|w^T(x_i^k-\overline{X}_k)\|^2.\end{aligned}\tag{6}$$

For simplicity of deriving the dual problem, we rearrange (6) into an equivalent quadratic programming (QP) problem, *as the optimization surrogate to approximate OrNCP*, and we call the QP problem QP-OrNCP and write it as

$$\begin{aligned}\min_{w,\xi}\quad & w^T\cdot S_w\cdot w+\lambda\cdot\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{p\neq k} w_{ikp}\cdot\xi_{ikp}\\ s.t.\quad & w^T(x_i^k-\overline{X}_l)\geqslant(k-l)-\xi_{ikl}\\ & \xi_{ikl}\geqslant 0\quad k=1,2,\ldots,K,\quad l=1,2,\ldots,k-1\\ & w^T(\overline{X}_h-x_i^k)\geqslant(h-k)-\xi_{ikh}\\ & \xi_{ikh}\geqslant 0\quad k=1,2,\ldots,K,\quad h=k+1,k+2,\ldots,K\end{aligned}\tag{7}$$

in which S_w represents within-class scatters matrix, w_{ikp}[2] denotes the ordinal weight defined jointly on the ith sample from the kth class and the class centroids of the kth and pth classes, λ is the trade-off parameter to control the model complexity, and ξ_{ikp} denotes the slack variables to regularize the solution space. It can be seen that the QP problem in (7) seeks for an optimal projection vector w along which (1) the within-class scatters are compacted maximally, i.e., the sum of distances of samples to their individual class centroids are minimized, as formulated in the objective of (7), and (2) the distances of samples from each class to the centroids of all the other classes are enlarged proportionally to corresponding class centroid–centroid discrepancies, as expressed in the constraints of (7). When taking the (1) and (2) together into account, the QP-OrNCP of (7) can be seen as a relaxation approximation to the OrNCP of (5). More importantly, compared with OrNCP of (5), QP-OrNCP can be more efficiently optimized.

As a standard QP problem, the QP-OrNCP of (7) can be solved by a variety of off-the-shelf optimization algorithms, such as *interior point* [21] and *conjugate gradient descent* [22], etc. On the other hand. we can also derive and optimize its dual problem instead, which is also a QP problem written as

$$\begin{aligned}\min_{\alpha}\quad & \frac{1}{4}\alpha^T\cdot\mathcal{A}^T\cdot S_w^{-1}\cdot\mathcal{A}\cdot\alpha-\mathcal{B}\cdot\alpha\\ s.t.\quad & 0\leqslant\alpha\leqslant\lambda\cdot vec(W)^T,\end{aligned}\tag{8}$$

where $\mathcal{A}:=[\overline{X}_2-x_1^1,\overline{X}_3-x_1^1,\ldots,x_{N_K}^K-\overline{X}_{K-1}]$, $\mathcal{B}:=[1,2,\ldots,1]$, [3] W is a tensor with the ikpth entry w_{ikp} being the ordinal weight defined as in Eq. (7), and $vec(\cdot)$ is the *vectorization operator*. Once the dual variable α has been obtained by optimizing Eq. (8), the primal ariable w can be calculated by

$$w=\frac{1}{2}S_w^{-1}\mathcal{A}\alpha.\tag{9}$$

With the projection vector w, the order of an unseen instance x can be predicted by the following decision rule[4]:

$$f(x)=\min_k\{k: w^Tx-b_k<0\},\tag{10}$$

where $b_k:=w^T(N_{k+1}\overline{X}_{k+1}+N_k\overline{X}_k)/(N_{k+1}+N_k)$.

3.3. Non-linear case of QP-OrNCP

To better handle the real-world possible non-linear OR problems, we can extend the QP-OrNCP to its non-linear counterpart by using kernel trick. In accordance with the Representer Theorem [26] associated with a feature mapping function $\phi: x\mapsto\phi(x)$, the original projection vector w can be represented by a combination of the training set as

$$w=\sum_{i=1}^{N}\beta_i\phi(x_i).\tag{11}$$

Substituting w by (11) into (7), we can derive the non-linear version of QP-OrNCP as

$$\begin{aligned}\min_{\beta,\xi}\quad & \beta^T\cdot\Theta\cdot\beta+\lambda\cdot\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{p\neq k} w_{ikp}\cdot\xi_{ikp}\\ s.t.\quad & \beta^T(Ker_i^k-\overline{Ker}_l)\geqslant(k-l)-\xi_{ikl}\\ & \xi_{ikl}\geqslant 0\quad k=1,2,\ldots,K,\quad l=1,2,\ldots,k-1\\ & \beta^T(\overline{Ker}_h-Ker_i^k)\geqslant(h-k)-\xi_{ikh}\\ & \xi_{ikh}\geqslant 0\quad k=1,2,\ldots,K,\quad h=k+1,k+2,\ldots,K\end{aligned}\tag{12}$$

where $\Theta:=\sum_{k=1}^{K}Ker_k(I-\mathbf{1}_{N_k})Ker_k^T$ with I being an identity matrix and $\mathbf{1}_{N_k}$ a matrix of all entries equaling $\frac{1}{N_k}$, Ker denotes the N-by-N kernel matrix with the ijth entry defined as $Ker_{ij}:=\phi(x_i)^T\cdot\phi(x_j)$, the N-by-N_k matrix Ker_k, corresponding to the kth class, is a sub-matrix of Ker, $\overline{Ker}_k:=[\frac{1}{N_k}\sum_{i=1}^{N_k}\phi(x_i^k)^T\cdot\phi(x_1),\ldots,\frac{1}{N_k}\sum_{i=1}^{N_k}\phi(x_i^k)^T\cdot\phi(x_N)]^T$ denotes the new centroid of the kth class in the mapped feature space, and $Ker_i^k:=[\phi(x_1)^T\cdot\phi(x_i^k),\ldots,\phi(x_N)^T\cdot\phi(x_i^k)]^T$ stands for the ith column of Ker_k. Through substituting the inner-product $\phi(\cdot)^T\cdot\phi(\cdot)$ defined in Eq. (12) with Mercer kernel functions such as *radial basis function* (RBF), we can map the samples from original feature space into a higher or even infinite-dimensional space. In form, the non-linear QP-OrNCP of (12) is analogous to its linear case of (7) and thus can be solved in a similar way.

[2] In this work, we set $w_{ikp}=|k-p|$. Actually, other reasonable definitions are acceptable as well.

[3] Note that the $\mathcal{A}$ and $\mathcal{B}$ in Eq. (8) are defined in order according to the left-hand and right-hand terms of the first and third lines of constraints in Eq. (7), respectively.

[4] It is worth to point out that for simplicity, the *nearest centroid rule* can also be used for decision making.

3.4. Complexity analysis of QP-OrNCP

In this subsection, we provide a complexity analysis to the QP-OrNCP. After analyzing the QP-OrNCP upon its primal and dual formulations, it can be found that solving the primal QP problem in (7) involves a $d \times d$ Hessian matrix associated with a total of $2N(K-1)$ linear constraints, where d denotes the feature dimension and N is the size of training set, while its dual problem in (8) involves a $[N(K-1)] \times [N(K-1)]$ Hessian matrix also with totally $2N(K-1)$ linear constraints. In our implementations, we adopt the interior-point algorithm [3] to solve the QP-OrNCP by optimizing its primal problem or dual form, respectively with a total complexity $\mathcal{O}(d^3)$ and $\mathcal{O}(N^3(K-1)^3)$, and it depends on the original feature dimension, the number of data classes, and the size of training set to whether solve the primal QP-OrNCP or its dual problem.

3.5. Relationship between KDLOR/MOR and QP-OrNCP

KDLOR vs QP-OrNCP: Comparing KDLOR of (1) with QP-OrNCP of (7), it can be proved that KDLOR actually is a reduced case of QP-OrNCP. Specifically, if we just consider two neighboring classes with equal ordinal weight $w_{ikp} = 1$ in the constraints of (7), then it degenerates into the following form

$$\begin{aligned} \min_{w,\xi} \quad & w^T \cdot S_w \cdot w + \lambda \cdot \sum_{k=2}^{K} \sum_{i=1}^{N_k} \xi_{ik} \\ s.t. \quad & w^T(x_i^k - \overline{X}_{k-1}) \geqslant 1 - \xi_{ik}, \quad k = 2,3,\ldots K, \quad i = 1,2,\ldots,N_k, \\ & \xi_{ik} \geqslant 0, \quad k = 2,3,\ldots K, \quad i = 1,2,\ldots,N_k. \end{aligned} \tag{13}$$

Summing up all the constraints associated with the N_k samples of the kth class and making an average, then Eq. (13) degenerates into

$$\begin{aligned} \min_{w,\sigma} \quad & w^T \cdot S_w \cdot w + \lambda \cdot \sum_{k=2}^{K} \sigma_k \\ s.t. \quad & w^T\left(\frac{1}{N_k}\sum_{i=1}^{N_k} x_i^k - \overline{X}_{k-1}\right) \geqslant 1 - \sigma_k, \quad k = 2,3,\ldots K, \\ & \sigma_k \geqslant 0, \quad k = 2,3,\ldots K, \end{aligned} \tag{14}$$

that is,

$$\begin{aligned} \min_{w,\sigma} \quad & w^T \cdot S_w \cdot w + \lambda \cdot \sum_{k=2}^{K} \sigma_k \\ s.t. \quad & w^T(\overline{X}_k - \overline{X}_{k-1}) \geqslant 1 - \sigma_k, \quad k = 2,3,\ldots K, \\ & \sigma_k \geqslant 0, \quad k = 2,3,\ldots K, \end{aligned} \tag{15}$$

where $\sigma_k := \frac{1}{N_k}\sum_{i=1}^{N_k} \xi_{ik}$. Till here, it can be found that (15) is essentially equivalent to KDLOR of (1) in strategy for OR, despite they differ in the scale of the thresholds. Therefore, KDLOR can be viewed as a reduced case of our QP-OrNCP.

MOR vs QP-OrNCP: Comparing (2) with (15), it can be found that if we define the S_w in (15) as the $X \cdot L \cdot X^T$ in (2), then MOR, in form, can also be regarded as a reduced case of QP-OrNCP, and the inference is similar to that for KDLOR above (from (13)–(15)).

Moreover, empirical comparisons between KDLOR/MOR and QP-OrNCP will be conducted in Section 4.

4. Experiments

To make evaluations on our proposed OR learning strategy, in this section we conduct a series of comparative experiments on synthetic and real-world datasets, respectively.

4.1. Experimental setup

In all experiments, we adopt the cross-validation technique for model selection. In non-linear cases, we uniformly adopt the RBF as mapping function, defined as $\phi(x_1)^T \cdot \phi(x_2) := exp\left(\frac{-(x_1-x_2)^2}{\delta^2}\right)$ where the bandwidth δ is tuned in the range of $\{0.01\overline{D}, 0.1\overline{D}, 0.2\overline{D}, \ldots, 0.9\overline{D}, \overline{D}, 10\overline{D}\}$ with $\overline{D} := \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{(x_i-x_j)^2}{N^2}$.

We apply the *Mean Absolute Error* (MAE) as the performance measure, defined as $MAE := \frac{1}{N}\sum_{i}^{N}(\hat{y}_i - y_i)$ with $\hat{y}_i$ and y_i denoting the predicted and ground-truth OR values, respectively.

4.2. Synthetic dataset

In order to make an intuitive comparison between our method (referring to the QP-OrNCP) and the KDLOR and MOR on their ability of performing OR, especially the ability to learn according to the distributions of data classes, in this subsection we conduct comparative experiment on a synthetic dataset.

For convenience of visual illustration, we generate a set of totally 4 ordinal classes, each with 20 two-dimensional examples sampled randomly (15 for training and the rest 5 for test), according to the specified *means* and *covariances* listed in Table 1. On the generated dataset, we respectively seek for the optimal OR projection direction using the KDLOR, MOR, and our method, and show the comparative results in Fig. 3.

As shown in Fig. 3, compared with the projection direction (represented by the line in orange-yellow[5]) trained using either the KDLOR or MOR, the direction (according to the line drawn in green) trained via our method, QP-OrNCP, can coincide more with the distribution trend of training data. Moreover, our method can yield a much smaller MAE of 0.3010 opposite to 0.3209 by the other two methods.

The experiment above shows the superiority of our method in performing OR, especially in capturing the distributions of data classes and taking such distribution information as side-information to optimize the solution space, compared with such typical methods like KDLOR and MOR.

4.3. Real-world datasets

To make further evaluation on the effectiveness of our strategy in handling real-world OR, in this section we conduct experiments on eight widely-used ordinal benchmark datasets,[6] hand-written digit recognition and human facial age estimation, respectively. Moreover, Besides the KDLOR and MOR, we also introduce the SVOR-IMX [7] and OHRank [6] into experiments for comparison due to their promising performance on OR. More importantly, to further investigate (1) the effectiveness of our strategy in jointly learning the thresholds and (2) the influence of granularity[7] of ordinal constraints on OR, we specially design two Foil methods for comparison, namely OR-M2M and OR-S2S, both of which share the same form of objective as QP-OrNCP of (7) but are imposed ordinal constraints, respectively just on class *centroids* and just on class *samples* (please refer to the Appendix for their detailed formulations).

[5] For interpretation of color in Fig. 3, the reader is referred to the web version of this article.

[6] available at: http://www.liacc.up.pt/~ltorgo/Regression/DataSets.html.

[7] The terminology "granularity" of data in this paper refers to such an unit for data partition, in which the *class sample* corresponds to a single data point, while the *class centroid* to an entire class, hence the granularity of the latter is larger than that of the former.

Table 1
The *means* and *covariances* of the four classes.

Class	Mean	Covariance
1	[0 0]	[100 0.1; 0.1 1]
2	[10 10]	[1 0.1; 0.1 100]
3	[20 20]	[100 0.1; 0.1 1]
4	[30 30]	[100 0.1; 0.1 100]

4.3.1. OR on ordinal benchmark datasets

Firstly, we make comparative experiments on the eight benchmark ordinal datasets to evaluate the effectiveness and superiority of our proposed OR learning strategy (i.e., the QP-OrNCP) in optimizing the thresholds in a joint way. Specifically, we group each of the 8 datasets into 10 equal-frequency bins (i.e., 10 classes), and report the comparative results over 10 random trials in Table 2.

It can be found from the results shown in Table 2 that among the 6 typical threshold-based OR methods, the proposed QP-OrNCP generates the lowest MAEs on the latter 5 of the totally 8 datasets. It is worth noting that in most cases, the QP-OrNCP defeats the SVOR-IMX which optimizes the thresholds automatically and jointly. It demonstrates the effectiveness and superiority of the proposed learning strategy in optimizing the ordinal thresholds on the eight benchmark datasets.

4.3.2. Hand-written digit recognition

We also conduct digit OR on two widely-used digit datasets[8] USPS and MNIST. Specifically, the USPS database contains a total of 11,000 samples for digits '0' to '9', each with 1100 samples. As for MNIST, it also consists of 10 digits of '0' to '9', and each digit has a quantity of 6313 samples. To exclude the interference caused by feature representation, we directly extract raw-pixels to represent the digit samples, and with consideration of mitigating the over-fitting, we uniformly normalize the digit images to 8×8 and the resulting feature dimension is 64. Examples from the USPS and MNIST are shown in Fig. 4(a) and (b), respectively.

For the hand-written digit recognition on the USPS and MNIST databases, we adopt 5-fold cross-validation for model selection, and show the experimental results averaged over 5 random runs in Figs. 5 and 6, respectively. From them, it can be found that,

1. Generally, MAEs of all the methods are decreasing with increasing training data. Moreover, in non-linear cases the performance difference among the methods is not so obvious as in linear cases, especially on USPS shown in Fig. 5(b). The reason can be found from the fact that MAEs of non-linear cases shown in Figs. 5(b) and 6(b) are correspondingly smaller than those of linear cases shown in 5(a) and 6(a), respectively, so the performance space is relatively smaller in non-linear feature space than in linear.
2. In most cases, the MAEs of OR-M2M are respectively smaller than those of KDLOR. By comparing (1) for KDLOR with (16) for OR-M2M, it can be found that the key difference between them relies on *the form of constraints*: in the OR-M2M the ordinal constraints are jointly imposed on all the class centroids (except for the difference in constraint granularity, OR-M2M takes the same form of ordinal constraints as our method QP-OrNCP), while in KDLOR the ordinality is imposed on a sequence of partial order constraints between neighboring two class centroids. Therefore, it demonstrates the superiority of jointly optimizing the ordinal thresholds.

[8] Note that the hand digit datasets may be used for cost-sensitive estimations, besides for OR [17].

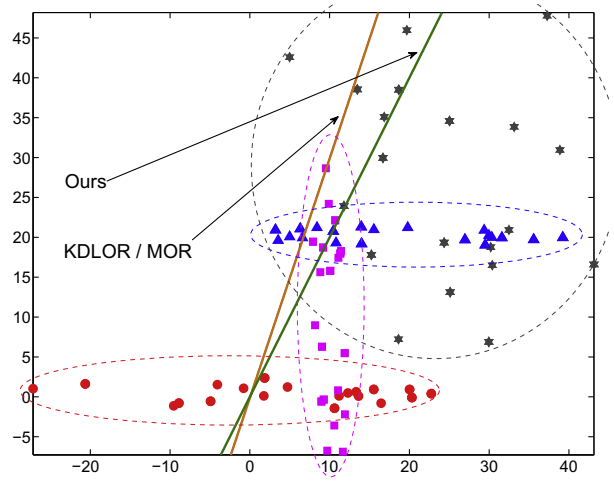


Fig. 3. Comparison of the projection directions trained by the KDLOR, MOR, and our method, respectively. Along their individual projection directions for OR test, our method QP-OrNCP yields a smaller MAE of 0.3010 opposite to 0.3209 by the KDLOR or MOR. Note that in this case, the projection direction of the KDLOR happens to coincide with that of the MOR.

3. In all cases, the MAEs of QP-OrNCP are respectively smaller than those of OR-M2M, which demonstrates that relatively coarse granularity of constraints unfavorably influences the OR performance and may be insufficient to preserve the ordinality of data. On the other hand, excessively fine granularity of ordinal constraints not only increases the computing complexity, but also likely to lead to over-fitting with poor OR performance as the OR-S2S which purely imposes ordinal constraints on the samples. Therefore, it demonstrates that the granularity of constraints is another crucial factor to OR.
4. Interestingly, the MAEs of the QP-OrNCP (the proposed method) are mostly lower than those of the SVOR-IMX which optimizes the thresholds in a joint and automatic way. It demonstrates the effectiveness, especially the superiority, of the QP-OrNCP in performing OR.

4.3.3. Human age estimation

Besides the above OR experiments, human age estimation is another typical real-world OR problem. So in this subsection we additionally conduct experiments on two well known aging datasets: the FG-NET and Morph.

In the experiment, to generate equal amount of samples for each age class, we respectively select a subset from the FG-NET and the Morph Album 1. More concretely, from FG-NET we randomly select 23 images for each age ranging from 0 to 19 years old, accounting for 20 age classes and some examples are shown in Fig. 7(a). As for Morph dataset, we randomly pick a number of 23 age classes from 16 years old to 38, each containing 31 images as shown in Fig. 7(b). Then, on the selected FG-NET and Morph sets, we uniformly crop the interested face regions from the raw images and normalize them to 16×16 pixels based on eye centers. With consideration of excluding the performance benefit brought by feature representation, we directly extract raw pixels as feature representation with dimension 256. Moreover, to mitigate the *small samples problem* as possible, we adopt the PCA [14] on the feature to extract 95 percent principle components as new feature representation with resulting dimension of 10 and 18, on the FG-NET and Morph, respectively.

On the generated aging datasets FG-NET and Morph, we perform age estimation using related OR methods, in which we uniformly adopt 2-fold cross-validation for optimal model

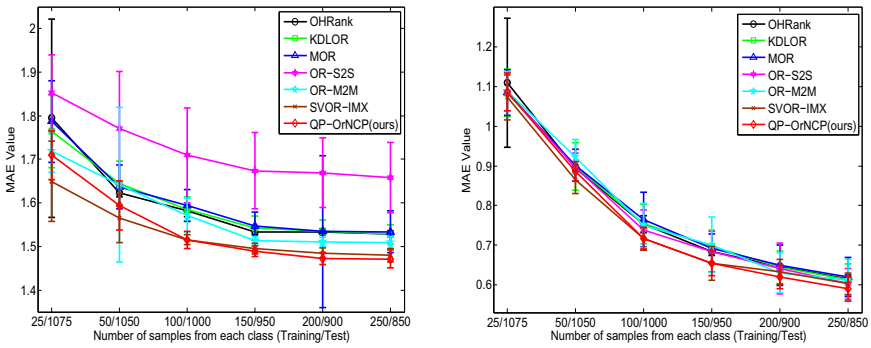
Table 2
Comparison (MAE ± STD) between different OR methods on the eight benchmark ordinal datasets.

Dataset	KDLOR	MOR	OR-M2M	OR-S2S	SVOR-IMX	QP-OrNCP
Pyrimidines	1.97 ± 0.09	**1.96 ± 0.08**	**1.96 ± 0.08**	2.00 ± 0.08	2.02 ± 0.10	1.98 ± 0.08
MachineCPU	**1.16 ± 0.07**	1.17 ± 0.05	1.17 ± 0.05	1.20 ± 0.04	1.17 ± 0.04	1.17 ± 0.06
Boston	1.15 ± 0.04	**1.13 ± 0.04**	1.14 ± 0.07	1.18 ± 0.04	1.14 ± 0.04	1.14 ± 0.06
Abalone	1.22 ± 0.06	1.20 ± 0.04	1.20 ± 0.10	1.26 ± 0.06	1.19 ± 0.03	**1.17 ± 0.05**
Bank	3.55 ± 0.03	3.52 ± 0.04	3.55 ± 0.04	–	3.58 ± 0.05	**3.50 ± 0.06**
Computer	1.23 ± 0.04	1.20 ± 0.03	1.15 ± 0.02	–	**1.08 ± 0.02**	**1.08 ± 0.01**
California	3.16 ± 0.02	3.15 ± 0.02	3.13 ± 0.02	–	3.14 ± 0.02	**3.10 ± 0.03**
Census	3.50 ± 0.03	3.50 ± 0.03	3.48 ± 0.05	–	3.47 ± 0.05	**3.41 ± 0.03**

Note: The "–" denotes that the results were not generated within 48 h in a computing platform with 16 GB memory and 4 CPU cores.
The results in bold are the best ones along each row.

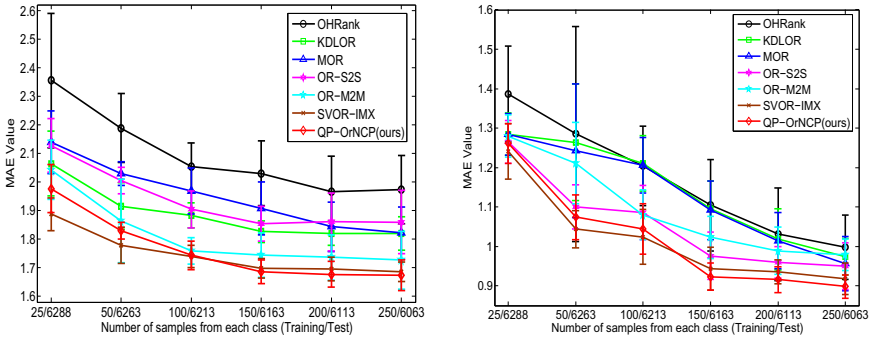
(a) USPS (b) MNIST

Fig. 4. Some examples from USPS and MNIST.



(a) Linear case (b) Non-linear case

Fig. 5. Comparison of digit OR results on USPS.



(a) Linear case (b) Non-linear case

Fig. 6. Comparison of digit OR results on MNIST.

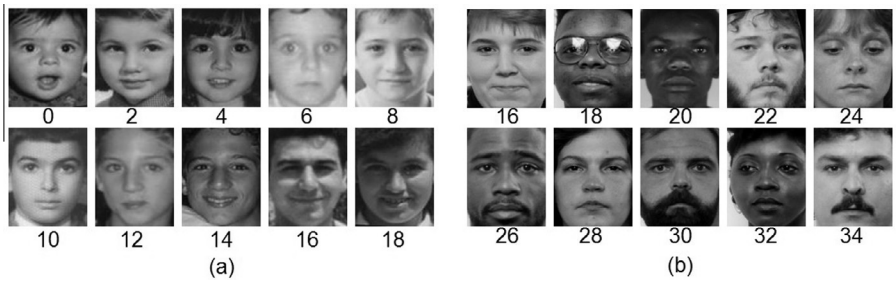
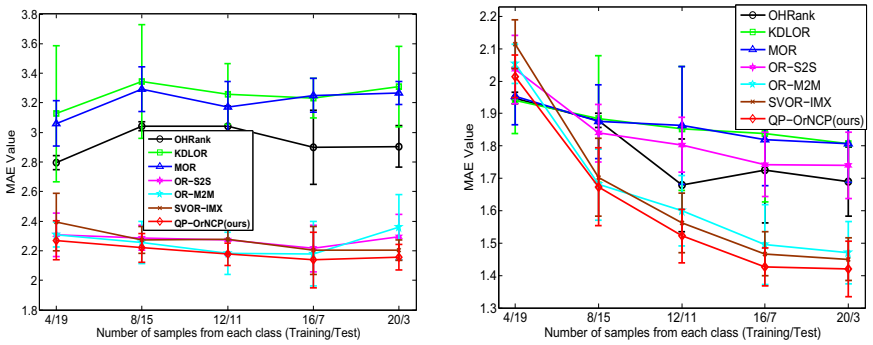
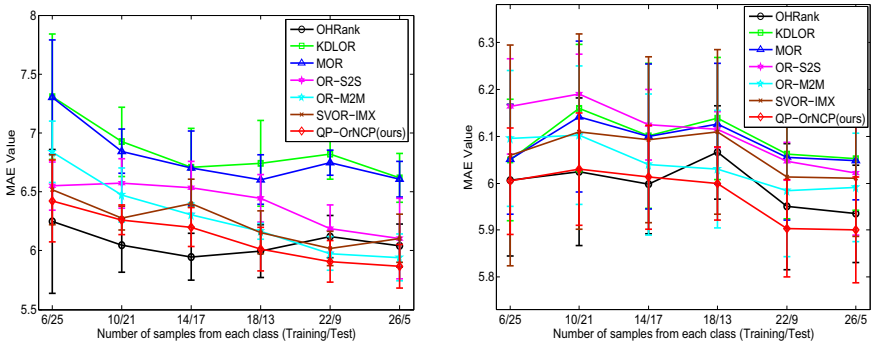


Fig. 7. Some normalized examples from the FG-NET (a) and Morph (b), where the number under each image denotes its corresponding facial age.



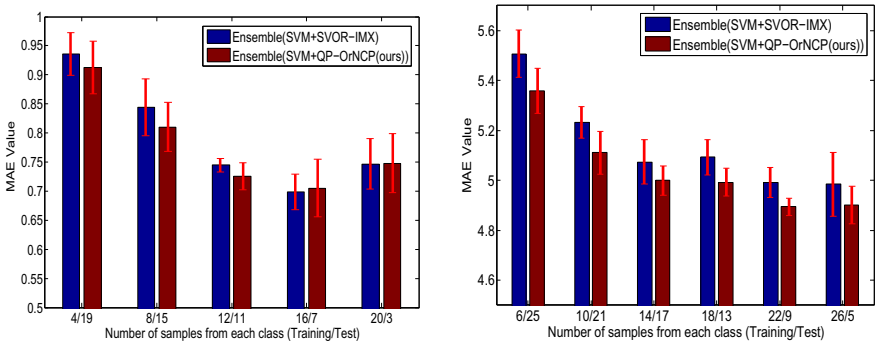
(a) Linear case (b) Non-linear case

Fig. 8. Comparison of human facial age OR results on FG-NET.



(a) Linear case (b) Non-linear case

Fig. 9. Comparison of human facial age OR results on Morph.



(a) On the FG-NET (b) On the Morph

Fig. 10. Comparison (MAE with the standard deviation) of ensemble learning on human age estimation. Here, the *Ensemble(SVM+SVOR-IMX)* stands for the ensemble method with the SVM and the SVOR-IMX as the binary and three ordinal classification methods, respectively.

selection. The experimental results averaged over 5 random runs are shown in Figs. 8 and 9, respectively. And it can be found from them that,

1. In all cases, the MAEs yielded by our method QP-OrNCP are correspondingly smaller than those by OR-M2M, OR-S2S, KDLOR and MOR. More importantly, compared with the state-of-the-art age estimation approach OHRank, our method QP-OrNCP can yield quite competitive estimation results, especially on the FG-NET. The results above prove the effectiveness and superiority of our learning strategy in performing ordinal age estimation.
2. The MAEs of the methods on Morph are correspondingly about 2 up to 3 times larger than those on FG-NET, which implies that the Morph dataset is relatively more difficult to identify than the FG-NET.

4.4. QP-OrNCP with ensemble learning

Following the success of ensemble learning in OR [25], we here consider to incorporate the QP-OrNCP as the ordinal classification method in ensemble learning for OR. Specially, without loss of generality, we respectively take the binary SVM and the QP-OrNCP as binary and three-class ordinal classification methods to construct our ensemble model, coined as *Ensemble(SVM+QP-OrNCP)*. To evaluate the effectiveness of the QP-OrNCP in promoting the ensemble learning, we make experimental comparisons on the relatively complex OR problem-human age estimation, and present the comparative results in Fig. 10.

By making a comparison between Figs. 10(a) and 8(a), and between Figs. 10(b) and 9(a), it can be found that:

1. The age estimation MAEs yielded by the ensemble methods are significantly lower than corresponding those by the standalone OR methods. It shows the superiority of ensemble strategy over the standalone.
2. The MAEs of the *Ensemble(SVM+QP-OrNCP)* are mostly lower than those of the *Ensemble(SVM+SVOR-IMX)*, especially on the Morph. It demonstrates that besides in the standalone OR learning, our proposed OR strategy, i.e., the QP-OrNCP, is also effective in promoting the ensemble based OR.

5. Conclusion

To perform better OR, in this work we proposed a novel strategy to jointly learn the thresholds that separate two neighboring ordinal classes across samples and class centroids through seeking for an optimal OR projection direction along which all the samples are distributed as in order as possible and simultaneously cater for a nearest-centroid distribution for each class. We call such a learning strategy *ordinal nearest-centroid projection learning* and for efficiency of optimization, we further relaxed it to a quadratic programming problem that in form covers the KDLOR and MOR as its special cases. Through extensive experiments on synthetic and real ordinal datasets, we demonstrated the effectiveness and superiority of our strategy in either standalone or ensemble OR learning.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant 61472186, the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20133218110032, the Funding of Jiangsu Innovation Program for Graduate Education under Grant *CXLX*13_159, and the Fundamental Research Funds for the Central Universities and Jiangsu *Qing-Lan Project*.

Appendix A

OR-M2M: Shares the same form of objective as QP-OrNCP of (7) but has global ordinal constraints imposed just on class *centroids*, formulated as

$$\begin{aligned}\min_{w,\xi}\quad & w^T\cdot S_w\cdot w+\lambda\cdot\sum_{k=1}^{K}\sum_{p\neq k}W_{kp}\cdot\xi_{kp}\\ s.t.\quad & w^T(\overline{X}_k-\overline{X}_l)\geqslant(k-l)-\xi_{kl}\\ & \xi_{kl}\geqslant 0\quad k=1,2,\ldots,K,\quad l=1,2,\ldots,k-1\\ & w^T(\overline{X}_h-\overline{X}_k)\geqslant(h-k)-\xi_{kh}\\ & \xi_{kh}\geqslant 0\quad k=1,2,\ldots,K,\quad h=k+1,k+2,\ldots,K\end{aligned}\tag{16}$$

where all the notations here have the same meanings as those in (7).

OR-S2S: Shares the same form of objective as QP-OrNCP of (7) but has global ordinal constraints imposed just on class *samples*, formulated as

$$\begin{aligned}\min_{w,\xi}\quad & w^T\cdot S_w\cdot w+\lambda\cdot\sum_{k=1}^{K}\sum_{p\neq k}\sum_{i=1}^{N_k}\sum_{j=1}^{N_p}W_{kpij}\cdot\xi_{kpij}\\ s.t.\quad & w^T(x_i^k-x_j^l)\geqslant(k-l)-\xi_{klij}\\ & \xi_{klij}\geqslant 0\quad k=1,2,\ldots,K,\quad l=1,2,\ldots,k-1\\ & w^T(x_j^h-x_i^k)\geqslant(h-k)-\xi_{khij}\\ & \xi_{khij}\geqslant 0\quad k=1,2,\ldots,K,\quad h=k+1,k+2,\ldots,K.\end{aligned}\tag{17}$$

The problems (16) and (17) are in form similar to (7), therefore, both can be solved by the same way for (7).

References

[1] G. Adomavicius, N. Manouselis, Y. Kwon, Multi-criteria recommender systems, in: Recommender Systems Handbook, Springer, 2011, pp. 769–803.
[2] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: NIPS, 2001, pp. 585–591.
[3] J.M. Borwein, A.S. Lewis, Convex Analysis and Nonlinear Optimization: Theory and Examples, vol. 3, Springer, 2010.
[4] J.S. Cardoso, J.F.P. Da Costa, Learning to classify ordinal data: the data replication method, J. Mach. Learn. Res. 8 (2007) 6.
[5] K.Y. Chang, C.S. Chen, Y.P. Hung, A ranking approach for human ages estimation based on face images, in: 2010 20th International Conference on Pattern Recognition (ICPR), IEEE, 2010, pp. 3396–3399.
[6] K.Y. Chang, C.S. Chen, Y.P. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 585–592.
[7] W. Chu, S.S. Keerthi, New approaches to support vector ordinal regression, in: Proceedings of the 22nd International Conference on Machine Learning, ACM, 2005, pp. 145–152.
[8] K. Crammer, Y. Singer, et al., Pranking with ranking, in: NIPS, 2001, pp. 641–647.
[9] E. Frank, M. Hall, A Simple Approach to Ordinal Classification, Springer, 2001.
[10] Q.B. Gao, Z.Z. Wang, Center-based nearest neighbor classifier, Pattern Recogn. 40 (2007) 346–349.
[11] C. Gentile, M.K. Warmuth, Linear hinge loss and average margin, in: NIPS, 1998, pp. 225–231.
[12] P.A. Gutiérrez, P. Tiňo, C. Hervás-Martínez, Ordinal regression neural networks based on concentric hyperspheres, Neural Networks 59 (2014) 51–60.
[13] T. Joachims, Optimizing search engines using clickthrough data, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 133–142.
[14] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2005.
[15] S. Kramer, G. Widmer, B. Pfahringer, M. De Groeve, Prediction of ordinal classes using regression trees, Fundam. Inform. 47 (2001) 1–13.
[16] H.T. Lin, L. Li, Reduction from cost-sensitive ordinal ranking to weighted binary classification, Neural Comput. 24 (2012) 1329–1367.
[17] Y. Liu, Y. Liu, K.C. Chan, Ordinal regression via manifold learning, in: AAAI, 2011.
[18] Y. Liu, Y. Liu, K.C. Chan, J. Zhang, Neighborhood preserving ordinal regression, in: Proceedings of the 4th International Conference on Internet Multimedia Computing and Service, ACM, 2012, pp. 119–122.
[19] M.J. Mathieson, Ordinal models for neural networks, Neural Networks Financ. Eng. (1996) 523–536.
[20] P. McCullagh, Regression models for ordinal data, J. Roy. Stat. Soc. Ser. B (Meth.) (1980) 109–142.

[21] Y. Nesterov, A. Nemirovskii, Y. Ye, Interior-point Polynomial Algorithms in Convex Programming, vol. 13, SIAM, 1994.
[22] J. Nocedal, S.J. Wright, Conjugate Gradient Methods, Springer, 2006.
[23] M. Pérez-Ortiz, P. Gutiérrez, C. Hervás-Martínez, Log-gamma distribution optimisation via maximum likelihood for ordered probability estimates, in: Hybrid Artificial Intelligence Systems, Springer, 2014, pp. 454–465.
[24] M. Pérez-Ortiz, P.A. Gutiérrez, M. Cruz-Ramírez, J. Sánchez-Monedero, C. Hervás-Martínez, Kernelizing the proportional odds model through the empirical kernel mapping, in: Advances in Computational Intelligence, Springer, 2013, pp. 270–279.
[25] M. Pérez-Ortiz, M. de la Paz-Marín, P. Gutiérrez, C. Hervás-Martínez, Classification of eu countries progress towards sustainable development based on ordinal regression techniques, Knowl.-Based Syst. 66 (2014) 178–189.
[26] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT press, 2002.
[27] A. Shashua, A. Levin, Ranking with large margin principle: two approaches, in: Advances in neural information processing systems, 2002, pp. 937–944.
[28] B.Y. Sun, J. Li, D.D. Wu, X.M. Zhang, W.B. Li, Kernel discriminant learning for ordinal regression, IEEE Trans. Knowl. Data Eng. 22 (2010) 906–910.
[29] B.Y. Sun, H.L. Wang, W.B. Li, H.J. Wang, J. Li, Z.Q. Du, Constructing and combining orthogonal projection vectors for ordinal regression, Neural Process. Lett. (2014) 1–17.
[30] Q. Tian, S. Chen, X. Tan, Comparative study among three strategies of incorporating spatial structures to ordinal image regression, Neurocomputing 136 (2014) 152–161.
[31] W. Waegeman, L. Boullart, An ensemble of weighted support vector machines for ordinal regression, Int. J. Comput. Syst. Sci. Eng. 3 (2009) 47–51.
[32] H. Wu, H. Lu, S. Ma, A practical SVM-based algorithm for ordinal regression in image retrieval, in: Proceedings of the Eleventh ACM International Conference on Multimedia, ACM, 2003, pp. 612–621.
[33] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, Multimodal classification of alzheimer's disease and mild cognitive impairment, Neuroimage 55 (2011) 856–867.