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

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).

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
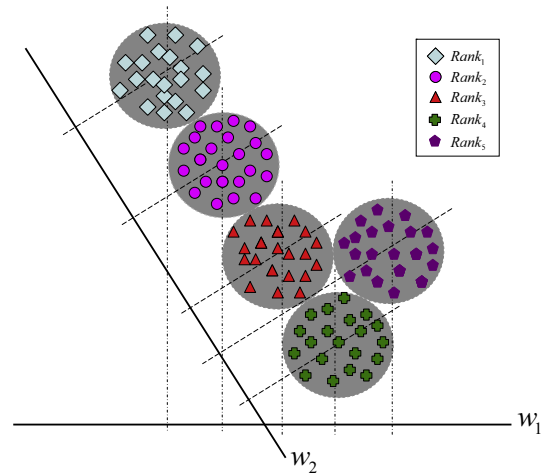

Fig. 1. Comparison between different OR projection directions, in which the direction of w_1 represents the one trained via MOR or KDLOR which encourages the OR learning to cater for their ordinal constraints built between class means of the data, whereas a more desirable direction may be the w_2 as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

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.

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

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

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.