Neurocomputing 388 (2020) 34–44

Contents lists available at ScienceDirect

Neurocomputing

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

Unimodal regularized neuron stick-breaking for ordinal classification

Xiaofeng Liu[a,b,1], Fangfang Fan[a,1], Lingsheng Kong[c], Zhihui Diao[c], Wanqing Xie[a], Jun Lu[a,*], Jane You[d]

[a] Beth Israel Deaconess Medical Center, Harvard Medical School, Harvard University, USA
[b] Department of ECE, Carnegie Mellon University, Pittsburgh, PA, USA
[c] CIOMP, Chinese Academy of Sciences, China
[d] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Article history:
Received 26 April 2019
Revised 14 September 2019
Accepted 9 January 2020
Available online 13 January 2020

Communicated by Dr. Bo Du

Keywords:
Ordinal regression
Deep neural network
Stick-breaking

ABSTRACT

This paper targets for the ordinal regression/classification, which objective is to learn a rule to predict labels from a discrete but ordered set. For instance, the classification for medical diagnosis usually involves inherently ordered labels corresponding to the level of health risk. Previous multi-task classifiers on ordinal data often use several binary classification branches to compute a series of cumulative probabilities. However, these cumulative probabilities are not guaranteed to be monotonically decreasing. It also introduces a large number of hyper-parameters to be fine-tuned manually. This paper aims to eliminate or at least largely reduce the effects of those problems. We propose a simple yet efficient way to rephrase the output layer of the conventional deep neural network. Besides, in order to alleviate the effects of label noise in ordinal datasets, we propose a unimodal label regularization strategy. It also explicitly encourages the class predictions to distribute on nearby classes of ground truth. We show that our methods lead to the state-of-the-art accuracy on the medical diagnose task (e.g., Diabetic Retinopathy and Ultrasound Breast dataset) as well as the face age prediction (e.g., Adience face and MORPH Album II) with very little additional cost.

1. Introduction

Recently, ordinal regression/classification has received much attention in recognition community. It aims to determine the discrete label of a certain pattern on an ordinal scale. The natural order of the labels (e.g., 1,2,3) indicates the order of the ranks [56].

For instance, the classes of a medical image usually represent the health risk levels, which are inherently ordered. The Diabetic Retinopathy Diagnosis (DR) involves five levels: no DR (1), mild DR (2), moderate DR (3), severe DR (4) and proliferative DR (5) [1]. The Breast Imaging-Reporting and Data System (BIRADS) also includes five diagnostic labels: 1-healthy, 2-benign, 3-probably benign, 4-may contain malignant and 5-probably contains malignant [2,3]. Similar ordinal labeling systems for liver (LIRADS), gynecology (GIRADS), colonography (CRADS) have been established soon afterward [4].

Surely, the ordinal data is not unique to the medical image classification [55]. Some other examples of ordinal labels include the age of a person [5,6], face expression intensity [7], aesthetic [8], star rating of a movie [9], etc., and are traditionally referred to ordinal regression tasks [10].

Recent advances in deep neural networks (DNN) for natural image tasks have prompted a surge of interest in adapting it to several applications [2,4,11]. However, some of the special characteristics of ordinal data have, in our opinion, not been efficiently explored.

Two of the most straightforward approaches for ordinal data either cast it as a multi-class classification problem [12] and optimize the cross-entropy (CE) loss or treat it as a metric regression problem [13] and minimize the absolute/squared error loss (i.e.,MAE/MSE). The former (Fig. 1(a)) assumes that the classes are independent of each other, which totally fails to explore the inherent ordering between the labels. The latter (Fig. 1(c)) treats the discrete labels as continuous numerical values, in which the adjacent classes are equally distant. This assumption violates the non-stationary property of many image related tasks, easily resulting in over-fitting [14].

Recently, better results were achieved via a $N-1$ binary classification sub-tasks (Fig. 1(b)) using sigmoid output with MSE loss [10] or softmax output with CE loss [3,4,15,16], when we have N levels as the class label. We can transform N levels to a series of

* Corresponding author at: Beth Israel Deaconess Medical Center, Harvard Medical School, Harvard University, USA
E-mail address: jlu@bidmc.harvard.edu (J. Lu).
[1] Xiaofeng Liu and Fangfang Fan contribute equally to this article.

https://doi.org/10.1016/j.neucom.2020.01.025
0925-2312/

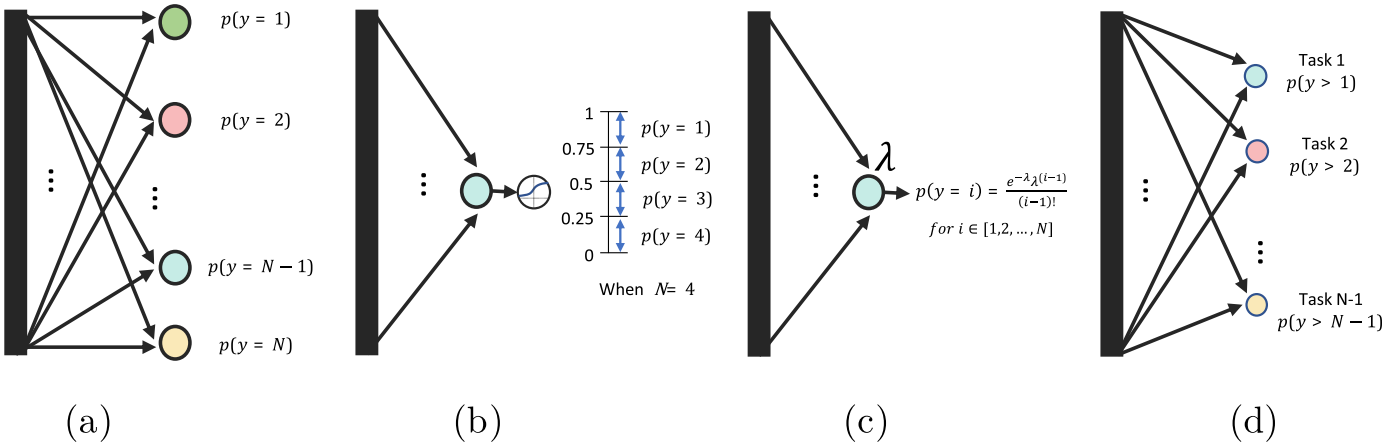

Fig. 1. The architecture of output layer used in previous ordinal regression methods: (a) multi-class classification, (b) regression, (c) Poisson, and (d) multi-task classification. We learn a discriminative mapping from sample x to an ordinal variable y.

labels of length $N-1$. Then the first class is [0, ... ,0], followed by the second class [1, ... ,0], third class [1,1, ... ,0] and so forth. The sub-branches in Fig. 1(b) calculate the cumulative probability $p(y > i|\mathbf{x})$, where i index the class.[1] With the cumulative probability, then it is trivial to define the corresponding discrete probabilities $p(y = i|\mathbf{x})$ via subtraction. These techniques are closely related to their non-deep counterparts [17,18]. However, the cumulative probabilities $p(y > 1|\mathbf{x}), \ldots, p(y > N-1|\mathbf{x})$ are calculated by several branches independently, therefore, can not guarantee they are monotonically decreasing. That leads to the $p(y = i|\mathbf{x})$ are not guaranteed to be strictly positive and results poor learning efficiency in the early stage of training. Moreover, $N-1$ weights need to be manually fine-tuned to balance the CE loss of each branch.

Besides, under the one-hot target label encoding, the CE loss $-\log(p(y = l|\mathbf{x}))$ essentially only cares about the ground truth class l. [19] argues that misclassifying an adult as a baby is more severe than misclassifying as a teenager, even if the probabilities of the adult class are the same. Authors in [20–22] propose to use a single output neuron to calculate a parameter of a unimodal distribution, and strictly require that the $p(y = i|\mathbf{x})$ follows a Poisson or Binomial distribution, but suffers from lacking the ability to control the variance [22]. Since the peak (also the mean and variance) of a Poisson distribution is equal to a designated λ, we can not assign the peak to the first or last class, and its variance is very high when we need the peak in the very later classes.

Furthermore, the quality of ordinal label makes this problem even more challenging. For example, the agreement rate of the radiologists for malignancy is usually less than 80% [23,24], which results in a noisy labeled dataset [25]. Despite the distinction between adjacent labels is often unclear, it is more likely that a well-trained annotator will mislabel a Severe DR (4) sample to Moderate DR (3) rather than No DR (1). The label smoothing is a general method for label noise, which cut down the 100% probability in one-hot distribution and averages it to all of the classes following a uniform distribution. It assumes the noise is caused by random error. However, there is a prior in medical diagnosis that a medical doctor or well-trained annotator is more likely to mislabel an image to a neighbor risk-level. Therefore, the uniform distribution may not be an ideal choice to model this kind of label noise. Instead, we propose to model it with the unimodal distribution. Besides, we show that smoothing the label with unimodal distribution can explicitly consider the relative similarity of ordinal data, and has a smaller loss when the prediction probabilities are closer distribute around the ground truth class.

In this paper, we propose to address the issues discussed above. The preliminary versions of the concepts in this paper were published in the BioImage Computing workshop at 2018 European Conference on Computer Vision [26]. In this paper, we extend those basic concepts in the following ways:

(1) We design a novel unimodal regularization strategy to smooth the target label from one-hot distribution to the unimodal distribution. It not only alleviates the effects of label noise in ordinal datasets, but also explicitly regularizes the structure of label space. Therefore, it can encourage the label predictions to distribute close to the ground truth class.
(2) We also investigate the possible combination of unimodal regularization with conventional models as well as the proposed neuron stick-breaking, which can improve the performance without sophisticated network design.
(3) We conduct all experiments using the new architecture, test on more general and challenging benchmarks, and give more comprehensive ablation studies.

In summary, this paper makes the following contributions.

(1) We rephrase the conventional softmax-based output layer to the neuron stick-breaking formulations to guarantee the cumulative probabilities are monotonically decreasing, and not need hyperparameters to balance the branches in muti-task learning framework.
(2) The unimodal label smoothing not only considers the prior knowledge in noisy ordinal data, but also explicitly includes a structured relationship between neighboring classes and penalizes less when the inter-class distances are smaller.
(3) Extensive evaluations in Diabetic Retinopathy, Ultrasound BIRADS, Adience face and MORPH Album II age datasets demonstrate that our method outperforms many state-of-art approaches on the medical diagnoses well as face age prediction task.

2. Related works

2.1. Ordinal regression

The conventional ordinal regression approaches can be classified to three classes, i.e., naive, binary decomposition and threshold methods [27–29]. Following the development of deep learning [30,31], several works have been proposed to target the

[1] We will always index probabilities from zero for the remainder of this paper.

ordinal data. Authors in [3,26] propose the multi-task learning framework. However, the percentages of each class are not guaranteed to be positive, which may hurt the training especially in the early stage. Besides, there are $N-1$ weights to balance the branches, which is a hard task for manually tunning [56]. Liu et al. [32] incorporate the metric learning for data relationship analysis. The ground metric of the earth mover's distance can also be used to describe the similarity of each class [19]. Different from these methods, we propose to use the stick-breaking process and adapt it to the neural network structure.

2.2. Stick-breaking process

The stick-breaking considering the problem of break a stick with length 1 to N segments. It is closely associated with the associated Bayesian non-parametric methods, e.g., Sethuraman [33] used it in constructive definitions of the Dirichlet process [34]. It is a subset of the random allocation processes [35] and a generalization of continuation ratio models [36]. Khan et al. [37] further proposed its parameterization for Latent Gaussian Models (LGMs). Inspired by these previous works, we propose to rephrase the output layer of neural network following the stick-breaking formulations and target for the ordinal data.

2.3. Unimodality of ordinal data

Authors in [21,22] propose to enforcing the prediction to be a Poisson distribution. In their parametric version, the output of the neural network is a single sigmoid unit, which represents the parameter λ in Poisson distribution [55]. However, require the output strictly following a specific distribution could be a strong assumption [56]. Besides, it is not easy to control the variance of the resulting Poisson distribution. Beckham and Pal [22] introduce an additional temperature parameter to control the variance, but results in more complicate hyper-parameter tuning. In here, we propose to use an exponential function following the softmax to flexible adjust the shape of target label distribution, and analysed the performance of Poisson, Binomial and exponential distribution.

3. Proposed methods

3.1. Neuron stick-breaking for ordinal regression

In the stick-breaking approach, we define a stick of unit length between [0,1], and sequentially break off parts of the stick [33]. The length of generated bits can represent the discrete probabilities for that class.

When we make a breaking manipulation, the stick will be separated to two parts with the length of $\sigma(\eta_1)$ and $1-\sigma(\eta_1)$, respectively. Their length can represent the probability of the two classes. Then, we further break the remaining part $1-\sigma(\eta_1)$ by defending how much of the percentage $\sigma(\eta_2)$ should be cut off in $1-\sigma(\eta_1)$. This will further generate two bits with the length of $\sigma(\eta_2)(1-\sigma(\eta_1))$ and leave $(1-\sigma(\eta_2))(1-\sigma(\eta_1))$. The $\sigma(\eta_1)$, $\sigma(\eta_2)(1-\sigma(\eta_1))$ and $(1-\sigma(\eta_2))(1-\sigma(\eta_1))$ can represent the probability of three classes respectively. Mathematically, to break a stick to N bits can be written as:

$$\begin{aligned} &p(y=1|\eta_1)=l_1=\sigma(\eta_1)\\ &p(y=j|\{\eta_n\}_{n=1}^{j})=l_j=\sigma(\eta_j)\prod_{i=1}^{j-1}(1-\sigma(\eta_i)) \quad j=2,3,\cdots N-1\\ &p(y=N|\{\eta_n\}_{n=1}^{N-1})=l_N=\prod_{i=1}^{N-1}(1-\sigma(\eta_{N-1})) \end{aligned} \tag{1}$$

where length of each bit can be used to formulate to the probability of each class $p(y)$. We note that the conventional stick-breaking processing in LGMs or Dirichlet process usually not care the $p(y=N|\{\eta_n\}_{n=1}^{N-1})$, but it can has clear meaning in ordinal problem. A simple case ($N=4$) is shown in Fig. 2, and it is appealing that only $N-1$ breaking manipulation can get N bits.

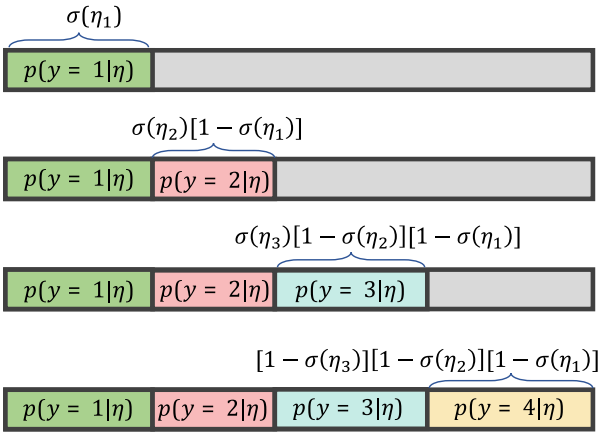

Fig. 2. The Stick-breaking process for 4 classes with 3 boundaries. In [37], η is the linear projection in LGMs.

To introduce the stick-breaking processes in a way that is appropriate a deep neural network for ordinal regression, we set $N-1$ output neurons for N levels as shown in Fig. 3. We suppose that $f(x)_i$ is a scalar denoting the i-th output of our neural network to substitute linear projections η_i in LGMs. We define the stick length of the first class, i.e., its probability, to be $\sigma(f(x)_1)$, where $\sigma(\cdot)$ denotes the sigmoid nonlinearity. We can then define the second class probability as what was left over from that stick multiplied by the output of the second class, i.e., $(1-\sigma(f(x)_1))\sigma(f(x)_2)$. For the third class probability we compute $(1-\sigma(f(x)_1))(1-\sigma(f(x)_2))\sigma(f(x)_3)$ and so forth, where the last class probability for $p(N|\mathbf{x})$ receives what is left over, i.e., $(1-\sigma(f(x)_1))...(1-\sigma(f(x)_{N-1}))$. The conventional CE loss can be used to train our network.

It can be derived that each output $f(x)_i$ is actually the ratio $f(x)_i=(p(y=i|x)/p(y\geq i|x))$, so these $f(x)_i$ can be interpreted as defining decision boundaries that try to separate the i-th class from all the classes that come after it. By doing so, the prediction is still a discrete probability (i.e., $\sum_i^{N=1} p(y=i)=1$), and each $p(y=i)\geq 0$, then we do guarantee the relationship of $p(y>1)\geq p(y>2)\geq p(y>N-1)$.

A nice property of our method is that unlike the approaches that only output a single distribution parameter [20,22,38], we obtain a slightly more expressive model since each boundary of two adjacent classes gets its own scalar output $f(x)_i$. The discrete probabilities can also be calculated via our predefined linear manipulations instead of having to estimate cumulative probabilities first [10,17,18]. Therefore, the weights of each branch in [10] are no longer necessary.

We repeatedly broke the part of the stick that remained in the stick-breaking process. We note that if we continue to break the first part of the stick, this is the Indian Buffet Process [39]. Therefore, the first part could be the probability of $p(y<N)$, and the length of the leaved part represent the $p(y=N)$. We also need $N-1$ break action to get the probability of N classes. Although the usage of them are different in LGMs, they are essentially the same in our applications.

3.2. Unimodal regularization

Label smoothing is a general regularization to address the noisy label problem, which encourages the model to be less confident

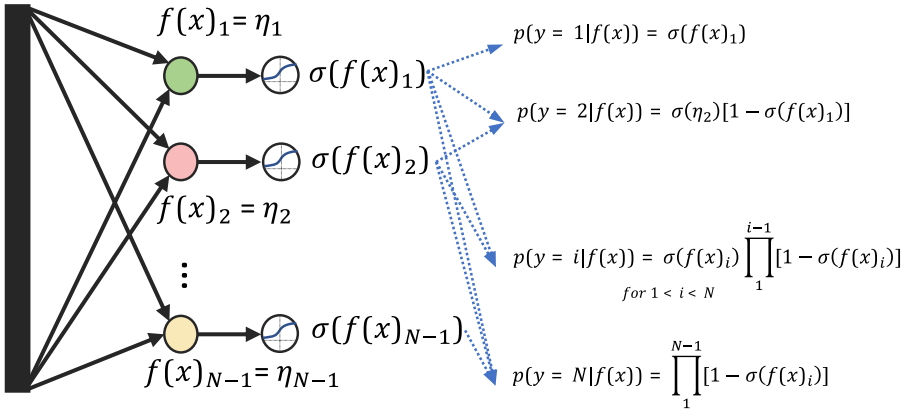

Fig. 3. Our neuron Stick-breaking architecture for N classes with $N-1$ output neurons, followed by sigmoid units and linear operations.

[40]. In the case of one-hot label, the distribution of a label probability is $q(i) = \delta_{i,l}$, where l is the ground truth class, $\delta_{i,l}$ is a Dirac delta, which equals to 1 for $i = l$, and 0 otherwise. The label smoothing replace $q(i)$ in CE loss (i.e., $\sum_i^N q(i)[-logp(y = i|)]$) with a more conservative target distribution [64].

$$q'(i) = (1 - \eta)\delta_{i,l} + \eta\frac{1}{N} \tag{2}$$

which can be regarded as the weighted sum of the original label distribution $q(i)$ and a fixed uniform distribution. Since ordinal data are more likely to be mislabeled as a class close to the true class, it is more reasonable to construct a group of unimodal distributions that have a peak at class l while decreasing its value when the class goes away from l. Another main requirement is that it can have a peak in the first as well as last classes, and the variances are expected to be low no matter the position of the peak.

There are three possible candidates are analyzed. The first is the Poisson Distribution. It is used to model the probability of the number of events, $k \in \mathbb{N} \cup 0$ occurring in a particular interval of time. Its probability mass function (PMF) is:

$$p_k = \frac{\lambda^k \exp(-\lambda)}{k!} \tag{3}$$

where $0 \le k \le K-1$, and $\lambda \in \mathbb{R}^+$ is the average frequency of these events. While we are not actually using it for the occurrence of an event, we can make use of its PMF to enforce discrete unimodal probability distributions. Its mean and variance are equal to the λ. As shown in Fig. 4, it is not easy to flexibly adjust the shape of a Poisson distribution. Besides, the probability of Poisson distribution for limited classes is not a distribution (i.e., the sum of each probability is not 1). To construct a distribution as our target label, we use a softmax function to normalize it.

Different from studies in [21,22] which predict the λ to define the shape of output distribution, we propose to modify the target label as a unimodal distribution. The output distribution does not strictly to be a unimodal distribution, but the cross-entropy loss will explicitly encourage it to form a unimodal distribution. Actually, the noise sample in ordinal data usually makes the unimodal to be a too strong assumption, while averaging some probability to every class following a unimodal distribution could achieve the similar function in a soft manner.

The second choice is Binomial Distribution. It is commonly adopted to model the probability of a given number of successes

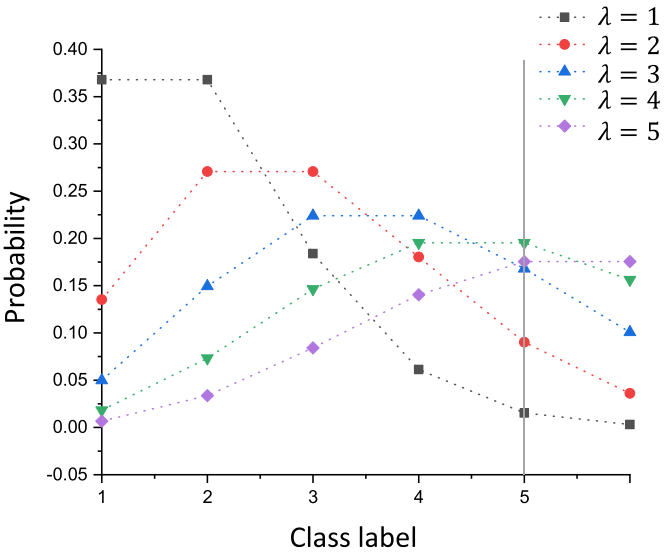

Fig. 4. The Poisson distribution-guided unimodal regularization for a dataset with 5 classes. We set *lambda* to the value of ground truth class and use p_k to present the probability of $k+1$ class. Then, followed by a softmax function to normalize the five class probabilities to a distribution.

out of a given number of trails k and the success probability p.

$$p_k = \binom{K}{k} p^k (1-p)^{K-k} \tag{4}$$

where $0 \le k \le K-1$. It has a mean of Kp and the variance of $Kp(1-p)$. A nice property of it is that we can define a range (e.g., $K = 5$) to distribute the probability and the sum of 5 probabilities is 1 as shown in Fig. 5. Therefore, we do not need the additional softmax function to normalize it. Even the mean (Kp) and the variance ($Kp(1-p)$) are different, it is still not easy to balance the position of its peak and the variance.

In here, we also propose to sample on an exponential function $e^{\frac{-|i-l|}{\tau}}$ and followed by a softmax normalization. Discrete distributions with 5 classes are illustrated in Fig. 6(a). Then, they are used to substitute the uniform distribution and construct the corresponding $q'(i)$. The loss is formulated as

$$\mathcal{L} = \sum_i^N q'(i)[-\log(p(y = i|\mathbf{x}))] \tag{5}$$

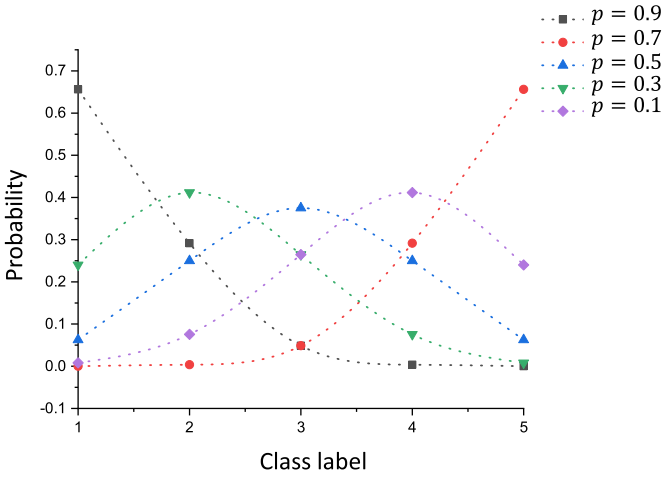

Fig. 5. The Binomial distribution-guided unimodal regularization for a dataset with 5 classes. We set K to the number of class (i.e., $K = 5$) and use p_k to present the probability of $k + 1$ class.

Since $q'(i)$ is monotonically decreasing w.r.t the farther distance from the true class l, we can regard it as a weight of $-\log(p(y = i|\mathbf{x}))$. From Fig. 6(b), this weight do explicitly consider the relative similarity of ordinal data, and has a smaller loss when the prediction probabilities are closer distribute around the l. Since the target label regularization can be processed advance, the training time does not increased by adding the unimodal regularization.

4. Numerical experiments

In this section, we show implementation details and experimental results on the Diabetic Retinopathy, Ultrasound BIRADS, Adience face and MORPH Album II age Datasets. To manifest the effectiveness of each setting choice and their combinations, we give a serial of elaborate ablation studies along with the standard measures. For a fair comparison, we choose similar backbones neural networks as in previous works. We adjust the last layer and softmax normalization to our neuron stick-breaking formulation. All of networks in our training use the $\mathcal{L}_2$ norm of 10^{-4}, ADAM optimizer [41] with 128 training batch-size and initial learning rate of 10^{-3}. The learning rate will be divided by ten when either the validation loss or the valid set QWK plateaus. The is no significant difference of the training time of NSB and multi-class classification, and the unimodal regularization is performed before the training stage.

4.1. Evaluations

There are several possible evaluation metrics for ordinal data [55,56]. As a classification problem, the performance of a system can be simply measured by the average classification accuracy. Ratner et al. [3] further utilized the Mean True Negative Rate (TNR) at True Positive Rate (TPR) of 0.95. The relatively high TPR used in here is fitted for strict TPR requirement of medical applications to avoid misdiagnosing diseased case as healthy. However, they do not consider the severity of different misclassification. Following the previous metrics in the Kaggle competition of DR dataset, we choose the quadratic weighted kappa (QWK)[2] to implicitly punish the misclassification proportional to the distance between the ground-of-truth label and predicted label of the network [42]. The QWK is formulated as:

$$k = 1 - \frac{\sum_{i,j} \mathbf{W}_{i,j}\mathbf{O}_{i,j}}{\sum_{i,j} \mathbf{W}_{i,j}\mathbf{E}_{i,j}} \tag{6}$$

to measures the level of disagreement between two raters ($\mathcal{A}$ and $\mathcal{B}$). In here, the $\mathcal{A}$ is the *argmax* prediction of our classifier and $\mathcal{B}$ is the ground truth. The $\mathbf{W}$ is a $N \times N$ matrix where $\mathbf{W}_{i,j}$ denotes the cost associated with misclassifying label i as label j. In QWK, $\mathbf{W}_{i,j} = (i - j)^2$. $\mathbf{O}_{i,j}$ counts the number of images that received a rating i by $\mathcal{A}$ and a rating j by $\mathcal{B}$. The quadratic calculation is one possible choice and one can plug in other distance metrics into kappa calculation. The matrix of expected ratings $\mathbf{E}$, is calculated, assuming that there is no correlation between rating scores. As a result, k is a scalar in $[-1,1]$, and $k = 1$ indicates the two raters are total agreement, whereas $k < 0$ means the classifier performs worse than random choice.

The Mean Absolute Error (MAE) metric is also popular in related ordinal datasets [10], which is computed using the average of the absolute errors between the ground truth and the estimated result. Here, we also propose its use in evaluating the proposed method on two medical ordinal benchmarks.

4.2. Diabetic Retinopathy (DR)

We make use of two typical ordinal datasets in the medical area suitable for DNN implementations. The first dataset contains images of Diabetic Retinopathy (DR).[3] In this dataset, a large amount of high-resolution fundus (i.e., interior surface at the back of the eye) images have been labeled as five levels of DR, with levels 1–5 representing the No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR, respectively. The left and right fundus image from 17563 patients are publicly available. The ResNet [43] style model with 11 ResBlocks as in [22] has been adopted for DR dataset. We use four stick-breaking neurons as our output structure and calculate the $p(y = i|\mathbf{x})$ via the predefined linear operations.

Following the setting in [22], we adopt the subject-independent ten-fold cross-validation, i.e., the validation set consisting of 10% of the patients is set aside. The images belonging to a patient will only appear in a single fold, in this way we can avoid contamination. The images are also preprocessed as in [20,22] and subsequently resized as 256 × 256 size images. Some examples can be found in Fig. 7.

We conduct our experiments with the evaluation metrics discussed earlier. The results in DR dataset are shown in Table 1. Several baseline methods are chosen for comparison, e.g., multi-class classification with CE loss (MC), regression with MSE loss (RG), Poisson distribution output with CE loss (Poisson), multi-class classification with squared earth mover's (Wasserstein) distance loss (EMD2), and multi-task network with a series of CE loss (MT). The RG is usually worse than MC, but appear to be competitive w.r.t. MAE, since RG optimizes similar metric MSE in its training stage. The Poisson gets the lowest results in most of evaluations due to its uncontrollable variance. The EMD2 and MT are more promising than MC as they consider ordinal information. By addressing their limitations, we achieve state-of-the-art performance in all of the evaluation tasks using the neuron stick-breaking (NSB) and unimodal label regularization (UR).

We also compared the performance of UR using Poisson (P), Binomial (B) and exponential (E) distribution. We set K to 5, which average the probability to all of the 5 classes. We set our hyper-parameters $\eta = 0.15$, $\tau = 1$. We note that the shape of normalized Poisson distribution is not easy to control and variation

[2] https://www.kaggle.com/c/diabetic-retinopathy-detection#evaluation.

[3] https://www.kaggle.com/c/diabetic-retinopathy-detection.

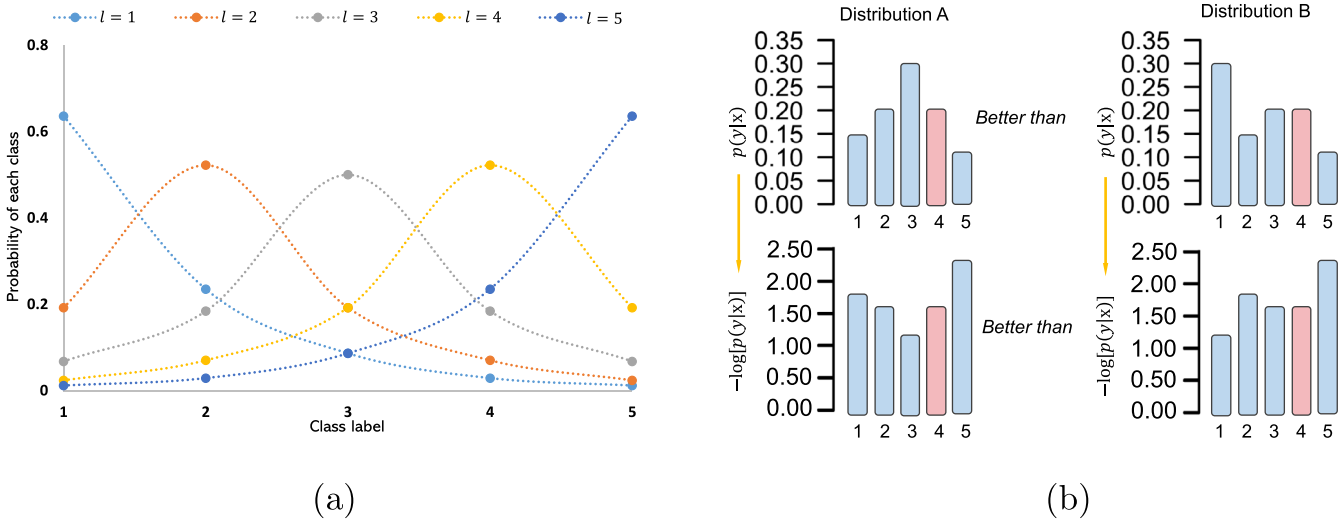

Fig. 6. (a) The distribution of normalized exponential function $e^{-|i-l|}$ for a dataset with 5 classes. (b) The expected (left) and inferior (right) distribution of predictions.

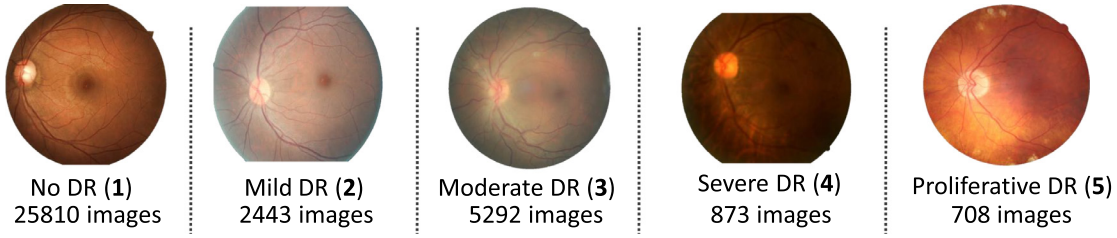

Fig. 7. Some samples with different retinopathy level in the DR dataset.

Table 1
Performance on the DR dataset. ± sd: standard deviation. The best performance are bolded.

Evaluations	Mean TNR@TPR=0.95			Valid Acc	Valid QWK	MAE
	1 *vs* 2-4	1-2 *vs* 3-4	1-3 *vs* 4			
MC	0.415	0.309	0.311	0.824	0.724	0.37
RG	0.403	0.306	0.308	0.762	0.705	0.38
Poisson [22]	0.388	0.300	0.296	0.771	0.713	0.38
EMD [19]	0.425	0.317	0.315	0.826	0.714	0.35
MT [3]	0.427	0.317	0.313	0.828	0.726	0.36
NSB	0.440 ± 0.02	0.331 ± 0.01	0.326 ± 0.02	0.842 ± 0.03	0.743 ± 0.03	0.32 ± 0.01
MC+UR(P)	0.418 ± 0.01	0.312 ± 0.01	0.315 ± 0.02	0.828 ± 0.03	0.728 ± 0.04	0.36 ± 0.00
MC+UR(B)	**0.422** ± 0.02	0.315 ± 0.02	0.318 ± 0.01	**0.830** ± 0.04	0.731 ± 0.03	**0.34** ± 0.01
MC+UR(E)	**0.422** ± 0.02	**0.316** ± 0.01	**0.319** ± 0.01	**0.830** ± 0.04	**0.732** ± 0.03	**0.34** ± 0.01
NSB+UR(P)	0.444 ± 0.01	0.332 ± 0.02	0.328 ± 0.01	0.844 ± 0.04	0.746 ± 0.04	0.31 ± 0.00
NSB+UR(B)	0.446 ± 0.02	**0.335** ± 0.02	**0.330** ± 0.01	0.846 ± 0.02	**0.748** ± 0.02	**0.29** ± 0.01
NSB+UR(E)	**0.447** ± 0.01	**0.335** ± 0.02	**0.330** ± 0.02	**0.847** ± 0.03	**0.748** ± 0.04	**0.29** ± 0.01

is different for each class. With more flexible shape, the Binomial and exponential case are more effective than Poisson distribution.

The UR contribute to consistent improvement. To analyze the impact of our hyper-parameters τ and η in UR(E), we performed some ablation studies. As shown in Fig. 8(a), the QWK is not sensitive to the $\tau \in \{0.8, 0.9, 1, 1.1\}$ when we fix the $\eta = 0.15$. Similarly, the QWK keep at the same level when we adjust η from 0.12 to 0.18 as shown in Fig. 8(b). We manually tune these hyperparameters with grid searching.

In Fig. 9, we also visualized the averaged prediction distribution on the test set of Diabetic Retinopathy. Although we only use the highest probability among all classes as our final prediction in the testing stage, the higher averaged probability indicates the higher confidence of the network for the corresponding class. It shows that the URNSB not only has the higher value in the ground truth class, but also distribute the probability closer the the ground truth label than the conventional multi-class classification.

4.3. Ultrasound BIRADS

The second medical dataset is the Ultrasound BIRADS (US-BIRADS) [3]. It is comprised of 4904 breast images which are labeled with the BIRADS system. Considering the relatively limited number of samples in level 5, we usually regard the 4–5 as a single level [3]. That results 2700 healthy (1) images, 1113 benign (2) images, 359 probably benign (3), and 732 may contain/contain malignant images. We divide this dataset into 5 subsets for

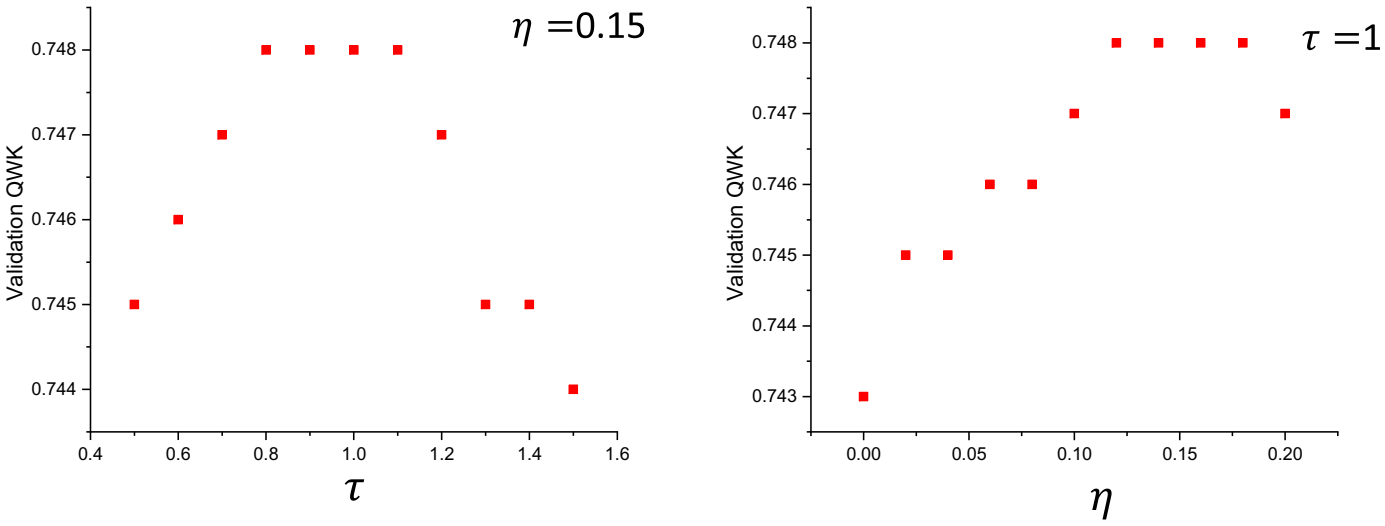

Fig. 8. The validation QWK is affected by the hyper-parameters: (a) τ (b) η.

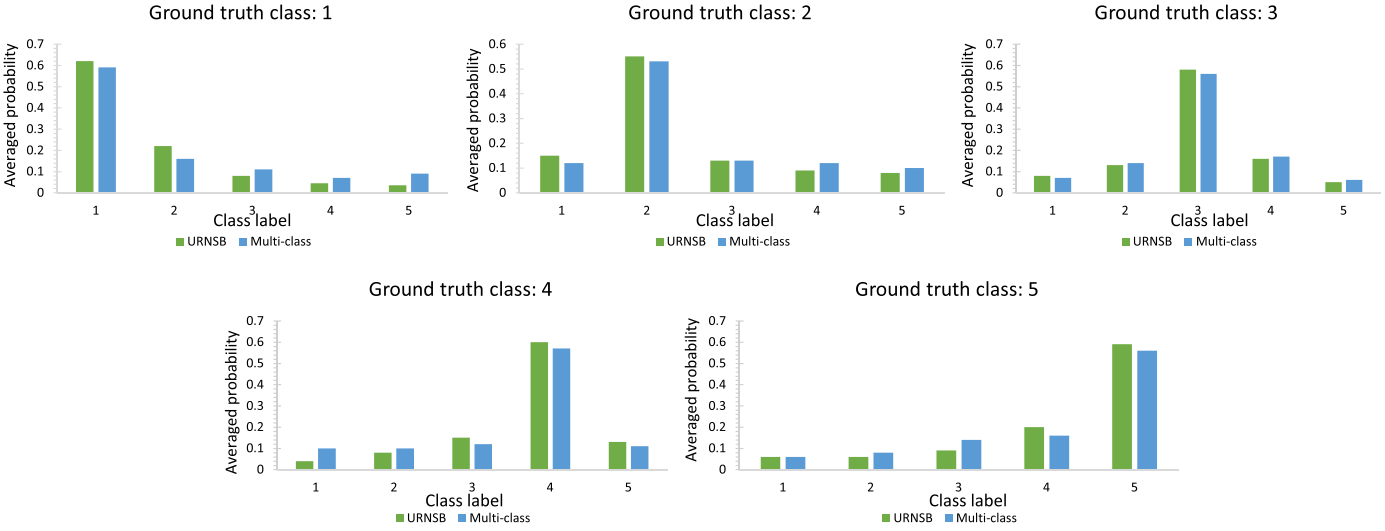

Fig. 9. The averaged prediction distribution on the test set of DR dataset. The NSB with exponential unimodal regularization (green) usually generates more concentrated output distribution around the ground truth class than conventional multi-class classification (blue).

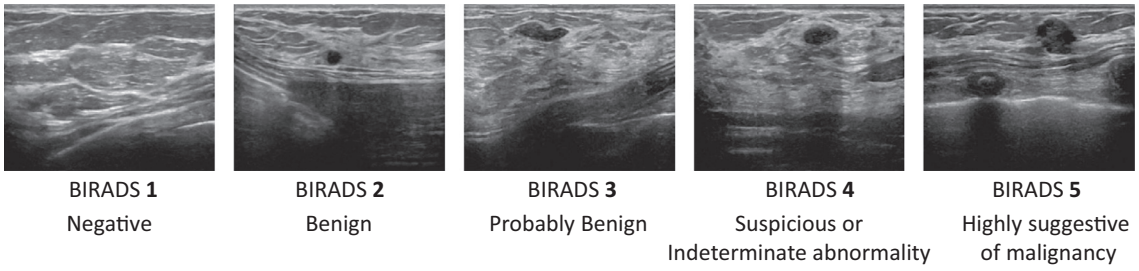

Fig. 10. Some samples with different malignant risk in the US-BIRADS.

subject-independent five-fold cross validation. We show some samples at different levels in Fig. 10.

AlexNet style architecture [44] with six convolution layers and following two dense layers is used for US-BIRADS image dataset as in [3]. 3 stick-breaking neurons are employed as the last layer. We set K to 5, $\eta = 0.15$, and $\tau = 1$.

The leading performance of our method is also observed on the US-BIRADS dataset (Table 2). Since its labels are more noisy (more severe annotator-dependent problem), the UR usually offers a more appealing contribution to the results. The Binomial and exponential distribution consistently outperforms the normalized Poisson distribution.

Table 2
Performance on the US-BIRADS dataset.*Our implementations have slightly higher TNR using MC baseline than the results reported in [3]. ± sd: standard deviation. The best performance are bolded.

Evaluations	Mean TNR@TPR=0.95			Valid Acc	Valid QWK	MAE
	1 *vs* 2-5	1-2 *vs* 3-5	1-3 *vs* 4-5			
MC	0.332*	0.287*	0.298*	0.733	0.678	0.42
RG	0.316	0.285	0.295	0.730	0.677	0.44
Poisson [22]	0.296	0.272	0.295	0.722	0.665	0.45
EMD [19]	0.378	0.287	0.317	0.761	0.686	0.41
MT [3]	0.385	0.292	0.313	0.765	0.685	0.41
NSB	0.391 ± 0.03	0.302 ± 0.02	0.320 ± 0.02	0.783 ± 0.03	0.694 ± 0.02	0.39 ± 0.01
NSB+UR(P)	0.395 ± 0.02	0.307 ± 0.02	0.324 ± 0.01	0.786 ± 0.02	0.697 ± 0.02	0.37 ± 0.00
NSB+UR(B)	0.396 ± 0.02	**0.309** ± 0.03	0.325 ± 0.01	**0.788** ± 0.03	0.699 ± 0.01	**0.36** ± 0.01
NSB+UR(E)	**0.397** ± 0.03	**0.309** ± 0.02	**0.327** ± 0.02	**0.788** ± 0.03	**0.700** ± 0.02	**0.36** ± 0.01

Fig. 11. Some examples from MORPH Album II Dataset.

4.4. MORPH Album II Dataset

Although our preliminary version [26] was originally developed for medical images, it is essentially applicable to other ordinal regression problems [57–63]. MORPH Album II [45] is one of the most popular benchmark for real age estimation. It contains 55,134 color images from 13,617 subjects with age and gender information. The age ranges from 16 to 77 years old. Some examples are give in Fig. 11, and we visualized the number of samples with each age label in Fig. 12. There are two testing protocol, the first is 5-fold random split (RS) and the second is the 5-folds subject-exclusive (SE) protocol.

Following [46], we choose the VGG-16 backbone and use 5-folds cross-validation. We use an initial learning rate of 0.001 and a batch size of 64 for VGG-16, and reduce the learning rate by multiplying 0.1 for every 15 epochs. We note that the age class is much more than the medical data. Therefore, we only smooth the probability to nearby classes instead of all of the classes. We choose the $K = 10$, $\eta = 0.15$, and $\tau = 1$. The relationship of K and the MAE is plotted in Fig. 13.

We note that when $K = 0$, it is equal to the NSB using one-hot target distribution. We see that the NSB and UR can generalize well in face image-based age estimation task (Tables 3 and 4).

4.5. Adience Face Age Dataset

We further test our method on Adience Face Age Dataset. It has 26580 images in 8 age groups from 2284 subjects. Some challenging sanples are shown in Fig. 14. Following [19], we choose a 40-layer residual network with identity mapping and

Table 3
Comparisons of the age estimation MAEs by the proposed approach and the state-of-the-art methods on the MORPH Album II. ± sd: standard deviation. The best performance are bolded.

Methods	MAE	Protocal
OR-CNN [10]	3.27	RS
DEX [47]	3.25	RS
DIF [48]	3.00	SE
Rank-CNN [16]	2.96	RS
RCL [49]	2.46/2.88	RS/SE
MVL [46]	2.4/2.80	RS/SE
Caps [50]	2.93 ± 0.05	SE
NSB	2.27 ± 0.03/2.69 ± 0.04	RS/SE
NSB+UR(P)	2.20 ± 0.02/2.64 ± 0.03	RS/SE
NSB+UR(B)	2.20 ± 0.03/**2.62** ± **0.04**	RS/SE
NSB+UR(E)	**2.19** ± **0.02**/**2.62** ± **0.03**	RS/SE

Table 4
Performance on the Adience Face Age dataset using conventional accuracy of exact match (AEM%) and with-in-one-category-off match (AEO%).

Methods	AEM	AEO
Dropout-SVM [51]	45.1%	79.5%
Cascade CNN [52]	52.9%	88.5%
EMD [19]	62.2%	94.3%
DEX [47]	55.6%	89.7%
DEX+IMDB+WIKI [47]	64.0%	96.6%
Gabor [53]	54.4	–
Caps [50]	59.8 ± 1.2%	–
NSB	65.7 ± 1.4%	95.8 ± 0.6%
NSB+UR(P)	66.1 ± 1.3%	96.0 ± 0.4%
NSB+UR(B)	**66.3** ± **1.5**%	96.4 ± 0.7%
NSB+UR(E)	**66.3** ± 1.2%	**96.5** ± **0.5%**

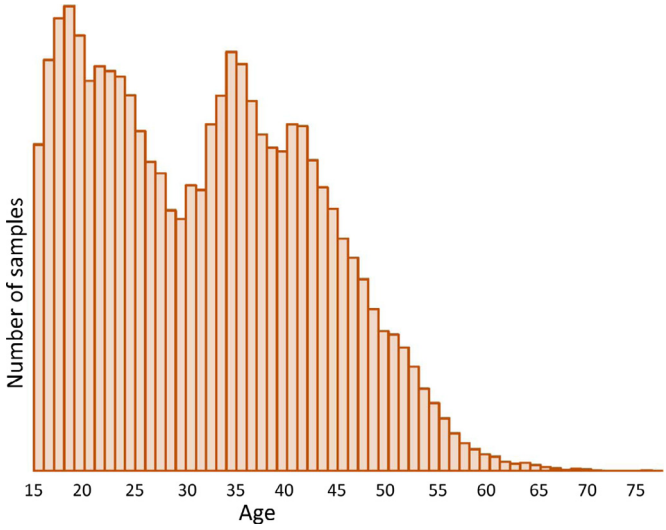

Fig. 12. The age distribution of MORPH Album II Dataset.

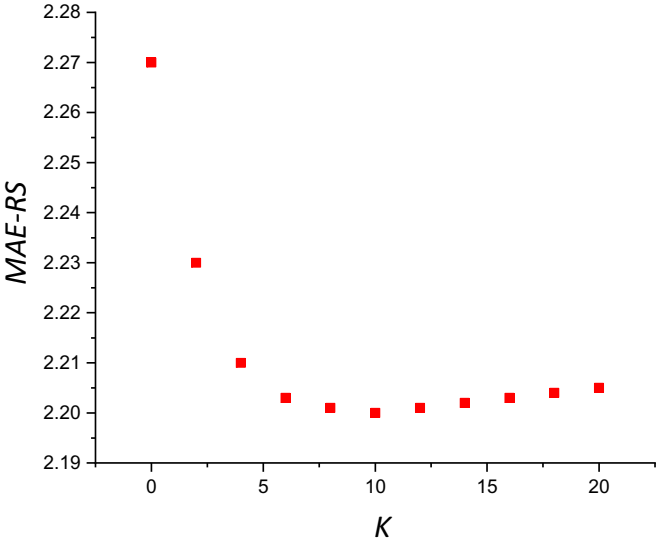

Fig. 13. The relationship of K and the MAE on the MORPH Album II Dataset.

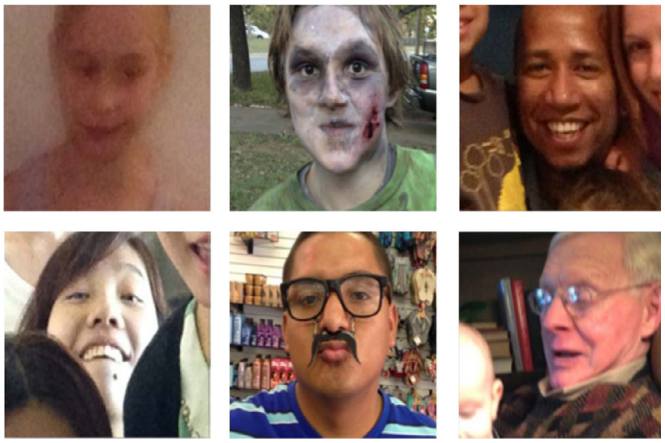

Fig. 14. Some challenging examples from the Adience Face Age dataset. ± sd: standard deviation. The best performance are bolded.

bottleneck design. We set batch size and learning rate to 128 and 10^{-2} respectively.

Then, we pre-train it on ImageNet. We set K to 8, $\eta = 0.15$, and $\tau = 1$. We see that our proposed methods outperform the previous methods again, which clearly shows that NSB and UR are effective and robust. Noticing that 524,230 additional images from IMDB and WIKI datasets are used in [47].

5. Conclusions

We have introduced the stick-breaking presses for DNN-based ordinal regression problem. By reformulating the neurons of the last layer and softmax function, we not only fully consider the ordinal property of the class labels, but also guarantee the cumulative probabilities are monotonically decreasing. Targeting on the noisy label problem in many datasets, we propose the unimodal label regularization, which has several attractive characteristics. We also show how these approaches offer improved performance in medical diagnose (*e.g.*, DR and US BIRADS) datasets and more general ordinal regression benchmarks (*e.g.*, Adience face and MORPH Album II Age Dataset). In future work, we intend to learn the hyperparameters in the UR to further reduce facilitate the manually tuning [54], and leverage our methods for more general ordinal regression tasks.

Declaration of Competing Interest

None.

Acknowledgment

This work was supported in part by the National Natural Science Foundation 61627819 and 61727818, Hong Kong Government General Research Fund GRF152202/14E, PolyU Central Research Grant G-YBJW, Youth Innovation Promotion Association, CAS (2017264), Innovative Foundation of CIOMP, CAS (Y586320150).

References

[1] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (22) (2016) 2402–2410.
[2] K.J. Geras, S. Wolfson, Y. Shen, S. Kim, L. Moy, K. Cho, High-resolution Breast Cancer Screening with Multi-view Deep Convolutional Neural Networks, arXiv preprint arXiv:1703.07047 (2017).
[3] V. Ratner, Y. Shoshan, T. Kachman, Learning Multiple Non-mutually-exclusive Tasks for Improved Classification of Inherently Ordered Labels, arXiv preprint arXiv:1805.11837 (2018).
[4] X. Li, Y. Kao, W. Shen, X. Li, G. Xie, Lung nodule malignancy prediction using multi-task convolutional neural network, in: Medical Imaging 2017: Computer-Aided Diagnosis, 10134, International Society for Optics and Photonics, 2017, p. 1013424.
[5] A.S. Al-Shannaq, L.A. Elrefaei, Comprehensive analysis of the literature for age estimation from facial images, IEEE Access 7 (2019) 93229–93249.
[6] R.R. Atallah, A. Kamsin, M.A. Ismail, S.A. Abdelrahman, S. Zerdoumi, Face recognition and age estimation implications of changes in facial features: a critical review study, IEEE Access 6 (2018) 28290–28304.
[7] R. Zhao, Q. Gan, S. Wang, Q. Ji, Facial expression intensity estimation using ordinal information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3466–3474.
[8] J.S. Cardoso, J.F.P. da Costa, M.J. Cardoso, Modelling ordinal relations with SVMs: an application to objective aesthetic evaluation of breast cancer conservative treatment, Neural Netw. 18 (5-6) (2005) 808–817.
[9] Y. Koren, J. Sill, Ordrec: an ordinal model for predicting personalized item rating distributions, in: Proceedings of the Fifth ACM Conference on Recommender Systems, ACM, 2011, pp. 117–124.
[10] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output cnn for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4920–4928.
[11] A.E. Gentry, C.K. Jackson-Cook, D.E. Lyon, K.J. Archer, Penalized ordinal regression methods for predicting stage of cancer in high-dimensional covariate spaces, Cancer Inform. 14 (2015) CIN–S17277.

[12] X. Geng, Z.-H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2234–2240.
[13] Y. Fu, T.S. Huang, Human age estimation with regression on discriminative aging manifold, IEEE Trans. Multimed. 10 (4) (2008) 578–584.
[14] K.-Y. Chang, C.-S. Chen, Y.-P. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 585–592.
[15] H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002–2011.
[16] S. Chen, C. Zhang, M. Dong, J. Le, M. Rao, Using ranking-cnn for age estimation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[17] J. Cheng, Z. Wang, G. Pollastri, A neural network approach to ordinal regression, in: Proceedings of IEEE International Joint Conference on Neural Networks, IJCNN (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 1279–1284.
[18] E. Frank, M. Hall, A simple approach to ordinal classification, in: Proceedings of European Conference on Machine Learning, Springer, 2001, pp. 145–156.
[19] L. Hou, C.-P. Yu, D. Samaras, Squared earth movers distance loss for training deep neural networks on ordered-classes, in: Proceedings of NIPS Workshop, 2017.
[20] C. Beckham, C. Pal, A Simple Squared-error Reformulation for Ordinal Classification, arXiv preprint arXiv:1612.00775 (2016).
[21] J.F.P. da Costa, H. Alonso, J.S. Cardoso, The unimodal model for the classification of ordinal data, Neural Netw. 21 (1) (2008) 78–91.
[22] C. Beckham, C. Pal, Unimodal Probability Distributions for Deep Ordinal Classification, arXiv preprint arXiv:1705.05278 (2017).
[23] R.M. Nishikawa, C.E. Comstock, M.N. Linver, G.M. Newstead, V. Sandhir, R.A. Schmidt, Agreement between radiologists interpretations of screening mammograms, in: Proceedings of International Workshop on Digital Mammography, Springer, 2016, pp. 3–10.
[24] A.J. Salazar, J.A. Romero, O.A. Bernal, A.P. Moreno, S.C. Velasco, Reliability of the BI-RADS final assessment categories and management recommendations in a telemammography context, J. Am. College Radiol. 14 (5) (2017) 686–692.
[25] B. Du, T. Xinyao, Z. Wang, L. Zhang, D. Tao, Robust graph-based semisupervised learning for noisy labeled data via maximum correntropy criterion, IEEE Trans. Cybern. 49 (4) (2018) 1440–1453.
[26] X. Liu, Y. Zou, Y. Song, C. Yang, J. You, B.V. Kumar, Ordinal regression with neuron stick-breaking for medical diagnosis, in: Proceedings of European Conference on Computer Vision, Springer, 2018, pp. 335–344.
[27] P.A. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, C. Hervas-Martinez, Ordinal regression methods: survey and experimental study, IEEE Trans. Knowl. Data Eng. 28 (1) (2016) 127–146.
[28] Z. Ma, S. Chen, A convex formulation for multiple ordinal output classification, Pattern Recognit. 86 (2019) 73–84.
[29] H. Zhao, Z. Wang, P. Liu, The ordinal relation preserving binary codes, Pattern Recognit. 48 (10) (2015) 3169–3179.
[30] X. Li, B. Du, Y. Zhang, C. Xu, D. Tao, Iterative privileged learning, IEEE Trans. Neural Netw. Learn. Syst. (2019) 1–13, doi:10.1109/TNNLS.2018.2889906.
[31] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, D. Tao, Stacked convolutional denoising auto-encoders for feature representation, IEEE Trans. Cybern. 47 (4) (2016) 1017–1027.
[32] Y. Liu, A.W.-K. Kong, C.K. Goh, Deep ordinal regression based on data relationship for small datasets, in: Proceedings of IJCAI, 2017, pp. 2372–2378.
[33] J. Sethuraman, A constructive definition of Dirichlet priors, Stat. Sin. 4 (1994) 639–650.
[34] B.A. Frigyik, A. Kapila, M.R. Gupta, Introduction to the Dirichlet Distribution and Related Processes, Department of Electrical Engineering, University of Washington, 2010. UWEETR-2010-0006.
[35] A. Agresti, An Introduction to Categorical Data Analysis, 135, Wiley, New York, 1996.
[36] P. Wan Kai, Continuation-ratio model for categorical data: a Gibbs sampling approach, in: Proceedings of the International MultiConference of Engineers and Computer Scientists, 1, 2008.
[37] M. Khan, S. Mohamed, B. Marlin, K. Murphy, A stick-breaking likelihood for categorical data analysis with latent gaussian models, in: Artificial Intelligence and Statistics, 2012, pp. 610–618.
[38] P.A. Gutiérrez, P. Tiňo, C. Hervás-Martínez, Ordinal regression neural networks based on concentric hyperspheres, Neural Netw. 59 (2014) 51–60.
[39] Y.W. Teh, D. Grür, Z. Ghahramani, Stick-breaking construction for the indian buffet process, in: Artificial Intelligence and Statistics, 2007, pp. 556–563.
[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[41] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980 (2014).
[42] J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, Psychol. Bull. 70 (4) (1968) 213.
[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[44] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[45] K. Ricanek, T. Tesafaye, Morph: a longitudinal image database of normal adult age-progression, in: Proceedings of 7th International Conference on Automatic Face and Gesture Recognition, FGR, IEEE, 2006, pp. 341–345.
[46] H. Pan, H. Han, S. Shan, X. Chen, Mean-variance loss for deep age estimation from a face, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5285–5294.
[47] R. Rothe, R. Timofte, L. Van Gool, Deep expectation of real and apparent age from a single image without facial landmarks, Int. J. Comput. Vis. 126 (2-4) (2018) 144–157.
[48] H. Han, C. Otto, X. Liu, A.K. Jain, Demographic estimation from face images: human vs. machine performance, IEEE Trans. Pattern Anal. Mach. Intell. 37 (6) (2015) 1148–1161.
[49] H. Pan, H. Han, S. Shan, X. Chen, Revised contrastive loss for robust age estimation from face, in: Proceedings of 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 3586–3591.
[50] S. Hosseini, N.I. Cho, Gf-capsnet: using Gabor jet and capsule networks for facial age, gender, and expression recognition, in: Proceedings of 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–8.
[51] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2170–2179.
[52] J.-C. Chen, A. Kumar, R. Ranjan, V.M. Patel, A. Alavi, R. Chellappa, A cascaded convolutional neural network for age estimation of unconstrained faces, in: Proceedings of 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE, 2016, pp. 1–8.
[53] H.J. Kwon, H.I. Koo, J.W. Soh, N.I. Cho, Age estimation using trainable Gabor wavelet layers in a convolutional neural network, in: Proceedings of 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 3626–3630.
[54] X. Liu, B.V. Kumar, P. Jia, J. You, Hard negative generation for identity-disentangled facial expression recognition, Pattern Recognit. 88 (2019) 1–12.
[55] X. Liu, Y. Zou, T. Che, P. Ding, P. Jia, J. You, B.V.K. Kumar, Conservative wasserstein training for pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8262–8272.
[56] X. Liu, X. Han, Y. Qiao, Y. Ge, S. Li, J. Lu, Unimodal-uniform constrained wasserstein training for medical diagnosis, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
[57] X. Liu, Y. Ge, C. Yang, P. Jia, Adaptive metric learning with deep neural networks for video-based facial expression recognition, Journal of Electronic Imaging 27 (1) (2018) 013022.
[58] X. Liu, L. Kong, Z. Diao, P. Jia, Line-scan system for continuous hand authentication, Optical Engineering 56 (3) (2017) 033106.
[59] X. Liu, S. Li, L. Kong, W. Xie, P. Jia, J. You, B.V.K. Kumar, Feature-Level Frankenstein: Eliminating Variations for Discriminative Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 637–646.
[60] X. Liu, Z. Guo, J. You, B.V. Kumar, Dependency-Aware Attention Control for Image Set-Based Face Recognition, IEEE Transactions on Information Forensics and Security 15 (2019) 1501–1512.
[61] X. Liu, Y. Zou, L. Kong, Z. Diao, J. Yan, J. Wang, … J. You, Data Augmentation via Latent Space Interpolation for Image Classification, in: 24th International Conference on Pattern Recognition, 2018, pp. 728–733.
[62] X. Liu, Z. Li, L. Kong, Z. Diao, J. Yan, Y. Zou, … J. You, A joint optimization framework of low-dimensional projection and collaborative representation for discriminative classification, in: 24th International Conference on Pattern Recognition, 2018, pp. 1493–1498.
[63] X. Liu, Z. Guo, S. Li, L. Kong, P. Jia, J. You, B.V.K. Kumar, Permutation-invariant feature restructuring for correlation-aware image set-based recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4986–4996.
[64] Y. Zou, Z. Yu, X. Liu, B.V.K. Kumar, J. Wang, Confidence regularized self-training, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5982–5991.

Xiaofeng Liu is a research fellow in Harvard Medical School, Harvard University. He was a joint supervision PhD in Carnegie Mellon University and University of Chinese Academy of Sciences. Before that, he received the B.Eng. degree in automation and B.A. degree in communication from the University of Science and Technology of China in 2014. He was the research assistant in MSRA and Facebook. He was a recipient of the Best Paper award of the IEEE International Conference on Identity, Security and Behavior Analysis 2018. His research interests include image processing, computer vision, and pattern recognition. He is the reviewer of CVPR, ICCV, ECCV, TPAMI, PR, etc.

Fangfang Fan is a research fellow at Harvard University. She received his Ph.D. degree from Huazhong University of Science and Technology in 2013. Her current research interests include emotion regulation and mental health as well as neural electrophysiology signal processing.

Lingsheng Kong received his bachelors degree from the University of Science and Technology of China in 2007 and his PhD in optical engineering from the University of the CAS in 2012. He is currently an associate Professor at CIOMP, CAS. His current research interests include optical engineering, imaging and image processing.

Zhihui Diao received his bachelors degree from Harbin Engineering University in 2010 and his PhD from the University of the CAS in 2015. He is currently an associate Professor with CIOMP, CAS. His current research interests include optical engineering, imaging and image processing.

Wanqing Xie a PhD major in Electronic Physics, is a visiting scholar at Harvard Medical School / Beth Israel Deaconess Medical Center. She also serves as an assistant professor at University of Science and Technology of China (USTC). She is interested in neuroscience, biomedical signal processing, bio-statistics, and machine learning.

Jun Lu is an Associate Professor of Neurology, Harvard Medical School. He received his MD from the Forth Military Medical University in 1984, MS from the Institute of Space Medico-Engineering in 1988, and PhD from Texas A M University in 1994.

Jane You received the Ph.D. degree from La Trobe University, Melbourne, VIC, Australia, in 1992. She is currently a Professor with the Department of Computing, Hong Kong Polytechnic University, Hong Kong, and the Chair of Department Research Committee. She has researched extensively in the fields of image processing, medical imaging, computer-aided diagnosis, and pattern recognition. She has been a Principal Investigator for one ITF project, three GRF projects, and many other joint grants since she joined PolyU in 1998. Prof. You was a recipient of three awards including Hong Kong Government Industrial Awards, the Special Prize and Gold Medal with Jurys Commendation at the 39th International Exhibition of Inventions of Geneva in 2011 for her current work on retinal imaging, and the Second Place in an International Competition [SPIE Medical Imaging2009 Retinopathy Online Challenge in (ROC2009)]. Her research output on retinal imaging has been successfully led to technology transfer with clinical applications. She is an Associate Editor of Pattern Recognition and other journals.