PGR-NET:A PARALLEL NETWORK BASED ON GROUP AND REGRESSION FOR AGE ESTIMATION

Na Liu, Liang Chang, Fuqing Duan *

College of Information Science and Technology, Beijing Normal University, Beijing 100875, China

ABSTRACT

Age is an important biometric feature of human face. Estimating the specific age of facial images is challenging, because of face aging's highly nonlinearity and randomness. Commonly age predictors are based on classification or regression method, which may be affected greatly by the category number or data distribution of the labelled samples. In this paper, we design a parallel deep neural network, called PGR-Net. It is a unified learning model which combines the merits of traditional classification methods and regression methods. The model consists of a classification network and several age regressors. The classification network is designed to divide facial images into several age groups, and a regressor is trained for each group separately. We train the classification network and the regressors in parallel, and perform age estimation with the regressor of the group predicted by the classification network. Experiments show that the proposed approach is fairly competitive compared with the state-of-the-art methods on two public datasets.

Index Terms— Age estimation, Classification, Regression

1. INTRODUCTION

Age is an important biological feature of human beings, and facial appearance will change greatly with the growth of age. Age information is very important for many applications in the field of human-computer interaction, and it also influences the performance of face recognition system. Age estimation from facial image is to model the changing of facial images with age, and predicts the approximate age of a person according to the facial image. However, appearance of facial images may be affected by head pose, illumination and facial expressions, and differs across gender, race and individual. So age prediction based on facial images is very challenging.

Age estimation is a special pattern recognition problem [3, 2, 11, 5, 6, 10, 7, 8, 13, 14, 20, 22] , which can be regarded as either a classification problem or a regression problem. In classification methods, Lanitis et al. [3] compared different classifiers such as quadratic functions, shortest distance classifier, supervised and unsupervised Neural Networks for automatic age estimation. Yang et al. [6] solved the ordinary binary classification problem using traditional LBPH feature. In recent years, many methods have been developed on designing deep learning classifiers. Levi et al. [2] proposed a simple convolutional net architecture to improve the classification accuracy. Chen et al. [10] took the ordinal relation between ages into consideration to get smaller classification errors. Rothe et al. [20] posed the age estimation problem as a deep classification problem. In regression methods, Fu et al. [7] applied quadratic regression on the discriminative aging manifolds. Guo et al. [8] designed a locally adjusted and robust regression. Lanitis et al. [14] used support vector regression for prediction of human age. More latter works mainly focused on regression methods. Shen et al. [13] proposed deep regression forests (DRFs) which divides soft data at split nodes to learn nonlinear regression. Xing et al. [22] compared three different methods and found the method based on regression achieved better performance.

In age classification, the classification accuracy depends highly on the category. When the total number of categories is relatively small, the classification accuracy will increase, but the accuracy of regression will decrease, and vice versa. Therefore, a good age classification method should balance both of the accuracy. In age regression, although it can predict the exact value of the age, the training of the regressor needs more dense data, and often is under-fitting. In this paper, we combine these two kinds of methods together and design an effective deep neural network to get better performance.

Our original idea is inspired by the work [1] in estimating image depth using any conventional monocular 2D camera, which used a multi-layer decision forests including a depth classification forest and a depth regression forest. It was a novel work performing classification and regression in order and get good results. But it relied too much on the depth of the forest whose determination was difficult. Later, Liu et al. [12] fused regression models of real-value with classification models based on Gaussian label distribution. Tan et al. [4] analyzed the relationship between the real age and its adjacent ages, and transformed the age estimation to $T + 3$ binary sub-problems by dividing the age into n groups, where T is the maximum age. In these methods, there are so many groups that the whole model is large and the implementation is complicated. Due to this, we adopt a simpler grouping method to implement a unified model combining the classification and regression.

The network in this work has an elaborated designed architecture with the combination of a classification network, a regression network and a judge layer. Traditional classification-based methods tend to achieve rough results due to the category number. Regression-based methods can achieve more accurate results, but they rely on large mount of dense and labelled data. Our method achieves a good balance and through a unified learning framework. Experimental results show that it can obtain fairly well results.

2. METHOD

2.1. PGR-Net

Since dividing ages into different groups is a piecewise and discrete function, the gradient of this kind of discontinuous process is difficult to calculate in the deep learning method. In general, it is not easy to implement if we want to integrate them into one network. Therefore, for simplicity of implementation, we design a PGR-Net.

PGR-Net is built from Alexnet. Alexnet consists of five convolution layers and three fully connected layers. The input images of 256×256 are cropped to 227×227, then fed into our network as input. Alexnet has lower accuracy than some deeper networks such

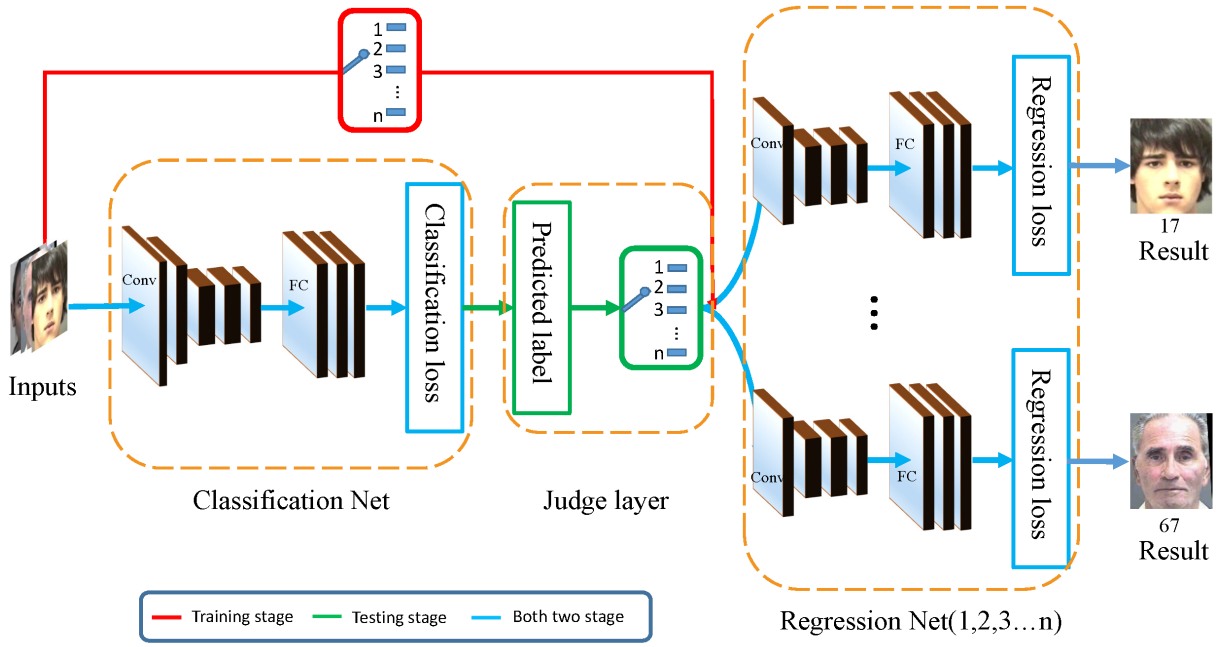

Fig. 1. PGR-Net: the main architecture of our network. red line and arrow is training stage, green represents test stage, blue is both training and testing stage. The classification and regression net is the basic Alexnet architecture, it consists of five convolutional layers and three fully connected layers. Red and green rectangles represent judge layers, they divide data and labels into n groups(blue strips in the rectangle).

as VGG and Res-Net, but its architecture is tiny and more flexible, which can shorten the training and testing time.

The network architecture is shown in Fig. 1. It consists of three parts, classification net, regression net and a judge layer, where the structures of the classification net and regression net are same to the structure of Alexnet. In the training stage, the architecture is a parallel net, the images and labels are input directly to the classification network. Meanwhile, The same data go through the judge layer, which distributes the data into several groups and regards each group data and corresponding label counted from zero as new input data into the regression network. In the testing stage, The cropped images pass the classification net firstly to obtain the predicted category label, and then the judge layer assigns the data into the corresponding regression network according to the predicted label. In the figure, we use the red and green lines to mark the components included in the training stage and the test stage respectively, and the blue line marks the component included in both two stages.

Judge layer in the figure is the key part. It integrates classification and regression networks into a unified learning model. Blue strips in the rectangle represent n classes divided from the classification net. Judge layer works as a switch, no matter which category the result of the classification network belongs to, the switch will route the result to corresponding regression network. The main function of this layer is splitting the data of corresponding groups and processing the labels while filling up batch simultaneously.

2.2. Loss Function

In classification net, we use the cross-entropy cost function[21], which is one of the most common classification loss functions, designed as the softmaxWithLoss layer in caffe.

For the orderly nonlinear problem of age estimation in regression net, ordinal regression [16] is very suitable. It transforms the age estimation problem into a series of simple binary sub-problems [26, 27], can be considered as a compromise between classification and regression. The predicted age $\hat{y}_j$ can be calculated as follows:

$$\hat{y}_j = 1 + \sum_{k=1}^{C-1} f_k\left(X_j\right)$$

where C represents the maximum of age, f_k is the k-th binary classifier, X_j is j-th test face.

We use the method in [15] to address ordinal regression problems and implement it as a new layer in Caffe.

2.3. Implementation Details

Parameter Setting. We implement the PGR-Net in the GPU mode with Caffe[19]. The network was trained with a weight decay of 0.0005 and a momentum of 0.9. The learning rate starts from 0.001 and is reduced by a factor of 10 along with the number of iterations increases. For all experiments, the network was initialized with the weights from training on ImageNet.

Batchsize Filling. In caffe, one batch data of a certain batchsize is processed at a time, and a network is only allowed to have one unique batch size during training and testing. After inputting, a batch of data will be divided into 5 groups by the judge layer and the number of segmented data must be less than or equal to batchsize. In order to keep the grouped data having the identical batchsize in each network to facilitate calculation, we need to design a optimal method to fill the segmented batch to original batchsize.

We have used many methods to conduct experimentssuch as picking a set of data to fill the batchsize randomly, or picking a set of best-performing data to fill. The first method has great randomness,

if we happen to choose a set of weakly performing data to fill, the result must be very poor, and vice versa. The second method is unfair to some extent. Filling with optimal data will definitely improve the experimental results, and it may masks the effect of our method. In summary, we duplicate the data cyclically to fill the batchsize and this loop filling strategy is of the most fairness.

For the case where no data is allocated to a certain group, the batchsize will be zero and there will be no data to perform the loop filling. Based on the existence of this case, we save and update one set of data in judge layer to solve this problem. If this happens, the saved data will be recalled to fill the batchsize. Due to the saved data are random, our method is fair to a certain extent.

3. EXPERIMENTS

3.1. Dataset

There are few datasets which can be used to estimate the age of 2D images. Two representative datasets are Morph Album2 [17] and Webface [25]. In the following, we will introduce them briefly.

Morph Album2. The Morph Album2 data contains 55,000 images of more than 13,000 people with different age, gender, and race. The age ranges from 16 to 77, and the average age is 33. Here we use the 5-fold cross-validation to evaluate the performance. The whole dataset is divided into five sets, four sets are used for training and the remaining one for testing. We don't adopt the segmentation method in [9], since [9] also considers the factors of gender and ethnicity.

Webface. WebFace dataset contains 62,203 images captured in the wild environment which involves large expressions and pose. The age ranges from 1 to 80, and the dataset is very challenging since it contains incomplete images and unreal facial images. We conduct experiments on Webface with a four-fold cross-validation [18].

3.2. Data Preprocessing and Evaluation Metric

3.2.1. Data Preprocessing.

Firstly, we align the face images. A face detector is used for recognition and an affine transformation is applied for registration. Five landmarks in face, i.e.left and right eyes corners, tip of the nose and two mouth corners, are selected as key points. All the aligned images are of the size 256×256. Then we feed the aligned images into the network.

Due to the complexity of Webface, we preprocess the dataset. Firstly, we use face++ to remove images with multiple faces; Then we reject images in different modalities such as face sketches and non-face images. Thirdly, we registrate the remaining images. The data preprocessing procedure of Webface is shown in Fig. 2.

3.2.2. Evaluation Metric

MAE. This paper uses the mean absolute error (MAE) to evaluate the regression result. MAE is calculated using the average of the absolute errors between the estimated value and the true value, i.e. $MAE = \frac{1}{m}\sum_{i=0}^{m-1}|y_i^{'} - y_i|$, where m is the number of testing face images, $y_i^{'}$ and y_i is the estimated value and the true value respectively.

Weighted Average All the regression results are weighted and averaged to get a final result. It can be calculated by function as follows:

$$f(x) = \min \sum_i w(i) \cdot f_i(x)$$

Table 1. Compared with different age interval in Morph folder-one

Age Interval	5	10	15	20	30
Number Of Groups	10	5	4	3	2
Mean Accuracy	0.652	0.795	0.827	0.913	0.921
Mean Mae	2.49	**2.29**	2.45	2.38	2.36

where x is the input images and i is the group we divide age into. w represents weights in every group, and $f_i(x)$ represent the optimal value obtained for each group in regression network.

We choose the softmax score as the final weights at first, and finally find that the experimental results are not satisfactory. Because if the data of younger age is assigned to an older group, the MAE is far from being offset by the weights. Besides, We choose the proportion of each group in the total data as the weights, but it is not reasonable in reality because the real grouping results of test dataset are often inaccurate and we don't know the true category. Therefore, it is more reasonable and effective to use the proportion of each group in the training sample as the weights to get the mean result directly.

3.3. Age Group and Upper Bound Analysis

Our method is based on the age classification first, so the key point is how many groups the age need to be divided into and how to divide them. We design three contrast experiments based on folder-one of Morph data which analysis the method of group selection. By conducting these experiments, we get the reliable group results. Finally, we divide the age into 5 groups with interval of 10, and we adopt adjacent grouping method.

Age Group Division. We need to find an optimal grouping so that the classification and regression can reach a balance. Too few groups will lead to the meaningless of PGR-Net, while too many groups will lead to a drop in classification accuracy. The original data were classified in the age interval of 5, 10, 15, 20, and 30 respectively. Table 1 shows the different results of different age interval. the first row represents the age interval, the second row is how many classes we get with the corresponding age interval. The third and final row represent the classification and regression results respectively. The figure shows that using 10 as the age interval is the best and ages are divided into 5 groups in Morph.

Adjacent Groups. Regression results in PGR-Net depend on the results of the classification, so a lower accuracy of classification can lead to a worse MAE. Therefore, in order to ensure the data a higher probability of being assigned to the right group, We merge adjacent groups into a new group. Experiments show that the adjacent groups is better than the original groups. In Table 2, The final column is the result of original groups, and the third column shows the result of adjacent groups. It can be inferred from the table that using the merge adjacent groups can get better results.

Upper Bound Analysis Suppose that the classification accuracy reaches 100%, the result of PGR-Net will be optimal. In this section, we denote the optimal value of our regression net to be an upper bound. That is, in regression net, we use the true label instead of the predicted results in classification.

We can see from the Table 2, the second column is the upper bound of our method. the result is reasonable and get a minimum which is better than other results obviously. Every efforts we make is to get the results as close to the upper bound as possible.

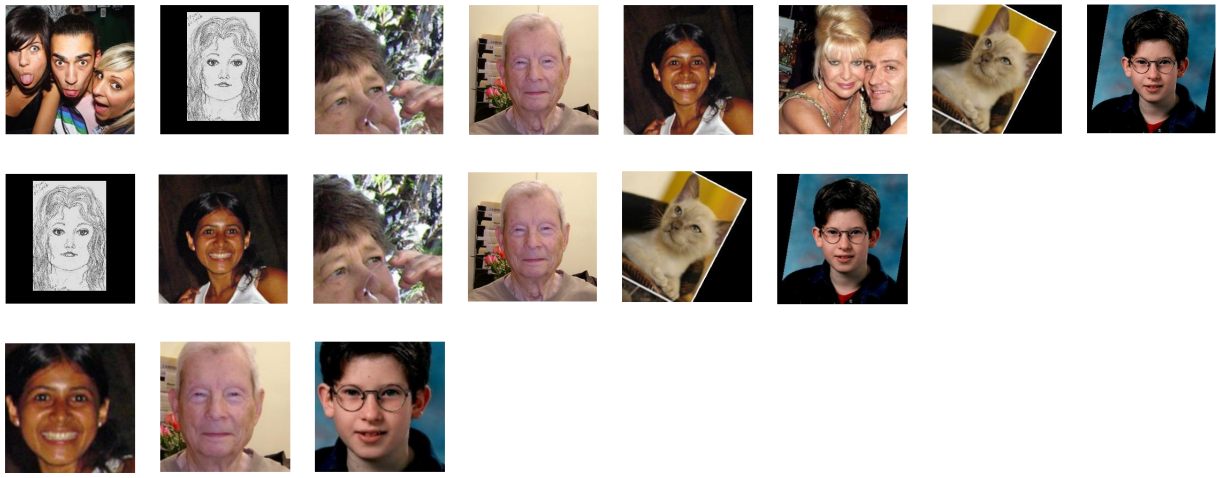

Fig. 2. The data preprocessing of Webface. The first row represents the original data (with the size of 62,203). It consists of frontal face images, multiple face images, partial face images, a small number of face sketch images, non-face images and etc. The second row represents the remaining images after deleting multi-facial images(with the size of 57478). The third row shows the remaining images after deleting sketch images and non-face images, then the remaining faces are aligned and then used in the experiments(with the size of 54203).

Table 2. Compared with diffenrent grouping method and the upper bound in Morph folder-one

	Upper Bound	Adjacent Groups	Original Groups
MAE	**2.095**	**2.292**	2.432

Table 3. Our experiment results in Morph and Webface

	fold 1	fold 2	fold3	fold 4	fold 5	mean
Morph	2.29	2.28	2.44	2.30	2.33	2.33
Webface	5.97	6.01	5.97	5.96	-	5.98

Table 4. The MAE value compared with state-of-the-art methods

	Morph	Webface
Tree-a-CNN[24]	-	7.72
DEX[20]	3.25	-
D2C[23]	3.06	6.04
DEL[22]	2.96	6.03
Ranking-CNN[10]	2.96	-
AGEn[4]	2.52	-
PGR-NET	**2.33**	**5.98**

3.4. Results

The experimental results using two datasets are shown in Table 3. We can see that the MAE of all the five folders does not change sharply except the folder3 whose MAE is 2.44, larger than other folders. A possible reason for it is uneven data distribution by checking the data. For Webface data, the results of all the four folders are uniform. The final result is the average of all the four folders. We can see that the final mean MAE is 2.33 and 5.98 for the two datasets respectively, and the result of Webface data is far worse than that of the Morph data. It is because the Webface data is captured in the wild environment, and is more challenging.

Table 4. shows the comparison with the state-of-the-art methods. We can see our PGR-Net achieves the best performance for both of the datasets. This shows that our method is competitive. In addition, we only need to finetune the original Alexnet model, while some other CNN-based methods need to pretrain the network with more datasets.

4. CONCLUSION

This paper proposes a unified parallel network called PGR-Net, which combines the merit of both classification and regression without any manual intervention. The model consists of a classification network and several age regressors. The classification network is designed to divide facial images into several age groups, and a regressor is trained for each group separately. We also prove the rationality of the parameters used in the experiment, such as the age interval we divided. Experiments show that the proposed approach is fairly competitive compared with the state-of-the-art methods on two public datasets.

In the future, there are also some improvements in our work. Firstly, if the basic network is changed to a deeper network such as VGG or Res-Net, the experimental results will be better. but the drawback is that deeper networks can lead to larger network architectures, which means higher demand for hardware configuration and lower training speed. So it's a trade-off between network architecture and the speed we need to make. Secondly, we should find the share layers between classify and regression network, which can reduce some parallel layers and simplify the models. Thirdly, the data processing can also be more elaborate, and the data should be evenly distributed to each fold as much as possible.

5. ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China under Grant No. 61572078, No. 61402040

6. REFERENCES

[1] Fanello S R, Keskin C, Izadi S, et al. Learning to be a depth camera for close-range human capture and interaction[J]. ACM Transactions on Graphics (TOG), 2014, 33(4): 86.

[2] Levi G, Hassner T. Age and gender classification using convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 34-42.

[3] Lanitis A, Draganova C, Christodoulou C. Comparing different classifiers for automatic age estimation[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2004, 34(1): 621-628.

[4] Tan Z, Wan J, Lei Z, et al. Efficient group-n encoding and decoding for facial age estimation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017 (1): 1-1.

[5] Wang C C, Su Y C, Hsu C T, et al. Bayesian age estimation on face images[C]//Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on. IEEE, 2009: 282-285.

[6] Yang Z, Ai H. Demographic classification with local binary patterns[C]//International Conference on Biometrics. Springer, Berlin, Heidelberg, 2007: 464-473.

[7] Fu Y, Huang T S. Human age estimation with regression on discriminative aging manifold[J]. IEEE Transactions on Multimedia, 2008, 10(4): 578-584.

[8] Guo G, Fu Y, Dyer C R, et al. Image-based human age estimation by manifold learning and locally adjusted robust regression[J]. IEEE Transactions on Image Processing, 2008, 17(7): 1178-1188.

[9] Guo G, Mu G. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression[C]//Computer vision and pattern recognition (cvpr), 2011 ieee conference on. IEEE, 2011: 657-664

[10] Chen S, Zhang C, Dong M, et al. Using ranking-cnn for age estimation[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[11] X. Geng, Z.-H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE Trans. Pattern Anal. Mach. Intel. 29 (12) (2007) 22342240.

[12] Liu X, Li S, Kan M, et al. Agenet: Deeply learned regressor and classifier for robust apparent age estimation[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015: 16-24.

[13] Shen W, Guo Y, Wang Y, et al. Deep Regression Forests for Age Estimation[J]. arXiv preprint arXiv:1712.07195, 2017.

[14] A. Lanitis, Comparative Evaluation of Automatic Age-Progression Methodologies, EURASIP J. Advances in Signal Processing,vol. 8, no. 2, pp. 1-10, Jan. 2008.

[15] Niu Z, Zhou M, Wang L, et al. Ordinal regression with multiple output cnn for age estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4920-4928.

[16] Herbrich R, Graepel T, Obermayer K. Support vector learning for ordinal regression[J]. 1999.

[17] Ricanek K, Tesafaye T. Morph: A longitudinal image database of normal adult age-progression[C]//Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on. IEEE, 2006: 341-345.

[18] ZHENG S. Visual image recognition system with object-level image representation[D]. , 2012.

[19] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[20] Rothe R, Timofte R, Van Gool L. Dex: Deep expectation of apparent age from a single image[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015: 10-15.

[21] Wong W S, Qin A. Method and apparatus for establishing topic word classes based on an entropy cost function to retrieve documents represented by the topic words: U.S. Patent 6,128,613[P]. 2000-10-3.

[22] Xing J, Li K, Hu W, et al. Diagnosing deep learning models for high accuracy age estimation from a single image[J]. Pattern Recognition, 2017, 66: 106-116.

[23] Li K, Xing J, Hu W, et al. D2C: Deep cumulatively and comparatively learning for human age estimation[J]. Pattern Recognition, 2017, 66: 95-105.

[24] Li S, Xing J, Niu Z, et al. Shape driven kernel adaptation in convolutional neural network for robust facial traits recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 222-230.

[25] Ni B, Song Z, Yan S. Web image mining towards universal age estimator[C]//Proceedings of the 17th ACM international conference on Multimedia. ACM, 2009: 85-94.

[26] Frank E, Hall M. A simple approach to ordinal classification[C]//European Conference on Machine Learning. Springer, Berlin, Heidelberg, 2001: 145-156.

[27] Li L, Lin H T. Ordinal regression by extended binary classification[C]//Advances in neural information processing systems. 2007: 865-872.