Ordinal Zero-Shot Learning

Zengwei Huo, Xin Geng*
MOE Key Laboratory of Computer Network and Information Integration,
School of Computer Science and Engineering,
Southeast University, Nanjing 210096, China
{huozw, xgeng}@seu.edu.cn

Abstract

Zero-shot learning predicts new class even if no training data is available for that class. The solution to conventional zero-shot learning usually depends on side information such as attribute or text corpora. But these side information is not easy to obtain or use. Fortunately in many classification tasks, the class labels are ordered, and therefore closely related to each other. This paper deals with zero-shot learning for ordinal classification. The key idea is using label relevance to expand supervision information from seen labels to unseen labels. The proposed method SIDL generates a supervision intensity distribution (SID) that contains each label's supervision intensity, and then learns a mapping from instance to SID. Experiments on two typical ordinal classification problems, i.e., head pose estimation and age estimation, show that SIDL performs significantly better than the compared regression methods. Furthermore, SIDL appears much more robust against the increase of unseen labels than other compared baselines.

1 Introduction

In some computer vision tasks, such as object classification, there are tens of thousands of different classes, only a few of which have been annotated. The class label space is huge and it is hard to have all classes available for training. In this case there are no training instances for some classes but we still want to make a prediction. Thus, the goal is learning to classify unseen class instances. This problem is generally referred to as zero-shot learning [Lampert *et al.*, 2014] [Palatucci *et al.*, 2009].

Most zero-shot learning methods are based on side information, such as attributes [Farhadi *et al.*, 2009] or text corpora [Socher *et al.*, 2013]. For example, [Palatucci *et al.*, 2009] uses the semantic knowledge bases to classify attributes; [Lampert *et al.*, 2014] proposes a probabilistic attribute prediction method; [Akata *et al.*, 2013] embeds each class into the space of attribute vectors and then learns a linear classifier; [Zhang and Saligrama, 2016] learns a joint latent space using structured learning.

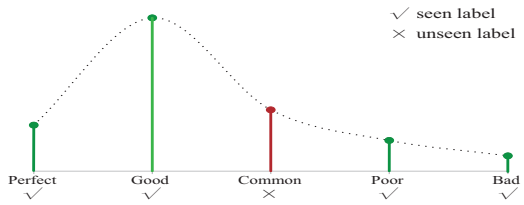


Figure 1: Supervision intensity for different labels. Green represents seen labels and red represents unseen labels. The ground truth label of this instance is "Good", so it has the strongest supervision intensity. Although "Common" is an unseen label, it still has certain supervision information because it is closely related to "Good".

The difficulty in obtaining the side information or using other techniques to process the side information are the most serious issues for many existing zero-shot learning methods. For the attribute-based methods, human experts are needed to label attributes and this is very time-consuming and not easy to obtain the discriminative category-level attributes. Some methods discover attributes interactively [Parikh and Grauman, 2011] [Branson *et al.*, 2010], but this also requires laborious human participation. Although many algorithms can discover attribute-related concepts on the Web [Rohrbach *et al.*, 2010] [Berg *et al.*, 2010], they can also be biased or lack information that is critical to a particular task [Parikh and Grauman, 2011]. For the text corpora-based methods, they first require a large language corpora, such as Wikipedia, and then need to learn word representation [Socher *et al.*, 2013] or use standard Natural Language Processing (NLP) techniques to produce class descriptions [Elhoseiny *et al.*, 2013]. It is hard to guarantee the correctness of such class descriptions for zero-shot learning. Conclusively, although side information is helpful for zero-shot learning, it has many disadvantages. Generating these side information is very tedious and sometimes we cannot know which side information is truly wanted. If we depend on human labor or NLP techniques, noisy side information will become almost inevitable and influence the final performance. To avoid these problems, it is important to solve zero-shot learning in whatever possible cases that have some properties we can utilize to avoid using

*Corresponding author.

side information.

One of the properties we can rely on is ordinal labels, which are quite common in real applications, such as facial beauty, student grades, restaurant evaluates, image scores and so on, where ordinal labels, such as "Perfect", "Good", "Common", "Poor" or "Bad", are used to represent levels, grades, degrees, etc. Ordinal zero-shot learning could be solved by standard regression methods if the ordinal labels are regarded as (sometimes a transformation process is needed) continuous real numbers. However, transforming ordinal labels into real numbers and vice versa often suffers certain information loss. Besides, it is hard to incorporate prior knowledge in the real numbers. Therefore, it is preferable for the ordinal classification problem to remain in the classification style, i.e., regarding the classes as discrete ordered labels rather than continuous numbers. Toward this goal, the relationship among the ordinal labels is used to expand supervision information from the seen classes to the unseen classes. Gradualness is perhaps the most common relationship among the ordinal label, i.e., once a label describes an instance, its neighboring labels in the order sequence may also describe the same instance to some extent less than the original label does. As shown in Figure 1, the ground truth label is "Good", so it has the strongest "supervision intensity", which means it describes the instance most accurately. The unseen neighboring label "Common" can also describe the instance to some extent due to its high relevance with "Good". It can be regarded that the "supervision intensity" of "Good" has been expanded to "Common". Different from conventional zero-shot learning, where the unseen labels offer no supervision information, in the method proposed in this paper, the supervision information of the unseen labels can be expanded from the seen labels according to the relationship among them. Combining the supervision intensity of all possible labels yields a data form similar to a probability distribution. We call it Supervision Intensity Distribution (SID) and propose an approach to learn effectively from SID for zero-shot learning problems.

The main contribution of this paper is to find a new solution for zero-shot learning with ordinal labels. Compared with previous methods, we use no side information and learn directly from the training data. To the best of our knowledge, this is one of the first attempts to solve the zero-short leaning problem without using any side information. Besides, we also study the unseen label patterns, which will influence the zero-shot learning performance. We conduct experiments in two computer vision tasks with ordinal labels: head pose estimation and age estimation. These experiments demonstrate that our method is effective and robust. The rest of the paper is organized as follows. Section 2 gives the definition of SID and introduces how to generate SID for a given instance, and then proposes the method to learn from SID. In Section 3, experimental results on head pose estimation and age estimation are reported. Finally, conclusions are drawn in Section 4.

2 Ordinal Zero-Shot Learning

In this section, we first introduce four patterns of unseen labels in ordinal label space, and then we give the definition of SID and introduce how to generate SID from ground truth label. Finally, we show the method of learning from SID.

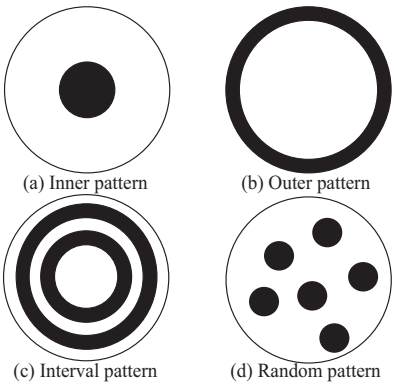


Figure 2: Four unseen label patterns in the ordinal label space. The disk represents the label space, and the dark parts represent the unseen labels.

2.1 Unseen Label Patterns

Previous works on zero-shot learning usually randomly choose unseen labels as test labels. This can also be used in ordinal zero-shot learning. But, since the label space is ordered and finite in ordinal classification, there are different patterns of unseen labels, which might influence zero-shot learning in different ways. Some patterns may be more challenging than the random pattern. The key idea of our method is expanding supervision information from seen labels to unseen labels. Different unseen label patterns may have different response to such expanding process. As can be seen in Figure 2, we study four typical unseen label patterns in this paper. The disk represents the entire ordinal label space and the dark parts represent the unseen labels. The four patterns are inner pattern, outer pattern, interval pattern and random pattern. For the inner pattern, the supervision information can only be expanded from the outside of the unseen labels; For the outer pattern, the supervision information can only be expanded from the seen labels surrounded by the unseen labels; For the interval pattern, the supervision information can be expanded from the seen labels between the unseen labels; Finally, for the random pattern, the expansion is also random. These four patterns are representative and fundamental and the combination of these patterns can generate more other patterns.

2.2 Supervision Intensity Distribution

For a given instance, each label in the label space can provide supervision information to this instance. We use supervision intensity to represent the strength of supervision information. Supervision intensity indicates the correlation of a label to an instance. The supervision intensity of all labels constitute a data form similar to a probability distribution which we call supervision intensity distribution (SID).

Assume the label space is $L = \{l_1, l_2, ..., l_C\}$, and given an instance $\boldsymbol{x}$, the SID is $\boldsymbol{y} = (y_1, y_2, ..., y_C)$, where C is the number of labels. The desired SID should satisfy two criteria. First is $y_i \in [0, 1]$ and $\sum_i y_i = 1$; Second is the ground truth label $\hat{l}$ is assigned with the highest supervision

intensity $\hat{y}$, and the supervision intensity decreases as the relevance between $\hat{l}$ and l_i decreases. That is, if the i-th label has the strongest correlation with $\hat{l}$, y_i will be higher than the supervision intensity of other labels less correlated with $\hat{l}$.

SID has a natural advantage for ordinal zero-shot learning. When learning for a particular label l, the instances assigned to other labels can also help because their labels are related to l according to the SID. This means that even if the training set has no instance with respect to a label, we can still use SID to learn this label.

However, the ground truth SID is not available in most datasets. Therefore, we need to transform the ground truth labels into SID before learning from SID. We use Mahalanobis distance to measure label relevance. When the labels are numbers, we can directly use the following formula to compute relevance between two labels,

$$r(l_i, l_j) = (l_i - l_j)^T \Sigma^{-1} (l_i - l_j), \tag{1}$$

$\mathbf{\Sigma}$ is a diagonal covariance matrix of $D \times D$, where D is the dimensionality of each label[1]. If class labels are not numbers, such as "Perfect", "Good", "Common", a straight-forward method is using evenly spaced integers to replace the labels[2]. Then, we can generate supervision intensity distribution by

$$y_i = \frac{1}{Z} \exp\left(-r(l_i, \hat{l})\right), \tag{2}$$

where Z is a normalization factor that ensures $\sum_i y_i = 1$, i.e.,

$$Z = \sum_i \exp\left(-r(l_i, \hat{l})\right). \tag{3}$$

Through transformation from the ground truth label to SID, the supervision information has been expanded to unseen labels from seen labels. Figure 3 gives one example. Assume there are 100 labels in the label space and these ordinal labels consist a square, we generate SID for each seen label outside the red box and the superposition of these SID consists a gray-scale image. Each pixel of the image corresponds to one label in the label space. For each pixel, higher intensity (brighter) means strong supervision information of the corresponding label. The labels inside the red box are unseen labels. Originally, they have no supervision information. But after expansion from the SID of the seen labels, the unseen labels inside the red box will also have supervision information of variant intensities. Note that the image in Figure 3 undergos a contrast stretching process to increase the image contrast for a better view.

2.3 Supervision Intensity Distribution Learning

After generating SID, the training set becomes $G = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, where $\boldsymbol{y}_i$ is the SID of instance $\boldsymbol{x}_i$, and y_i^j is the j-th element in $\boldsymbol{y}_i$. The goal is to learn a model that can predict a SID which is similar to $\boldsymbol{y}_i$ given the instance $\boldsymbol{x}_i$. This problem setting matches a recently proposed machine learning paradigm called Label Distribution Learning [Geng, 2016], where SID can be viewed as a special form of label distribution. There are many criteria to measure the similarity between two distributions and here we use Jeffrey's divergence. Jeffrey's divergence between two distribution $\boldsymbol{P}$ and $\boldsymbol{Q}$ is defined by

$$D_J(\boldsymbol{P}||\boldsymbol{Q}) = \sum_i (P_i - Q_i) \log \frac{P_i}{Q_i}, \tag{4}$$

where P_i and Q_i are the i-th element of $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively. As Geng described in [Geng *et al.*, 2013] and [Geng and Xia, 2014], we use the maximum entropy model to embody the mapping from the instance to its corresponding SID, i.e.,

$$p(l_j|\boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{1}{\Gamma_i} \exp(\sum_r \theta_{jr} \boldsymbol{x}_i^r), \tag{5}$$

where $\Gamma_i = \sum_k \exp(\sum_r \theta_{kr} \boldsymbol{x}_i^r)$ is the normalization factor, $\boldsymbol{x}_i^r$ is the r-th feature of $\boldsymbol{x}_i$, and θ_{jr} is an element in $\boldsymbol{\theta}$ corresponding to the label l_j and the r-th feature.

Then the best parameter $\boldsymbol{\theta}^*$ is determined by

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_i D_J(\boldsymbol{y}_i||p(\boldsymbol{l}|\boldsymbol{x}_i; \boldsymbol{\theta})) + \lambda_1 \sum_{(m,n)\in N} ||\boldsymbol{\theta}_m - \boldsymbol{\theta}_n||_2^2, \tag{6}$$

The first term is Jeffrey's divergence and $\boldsymbol{y}_i$ is the ground truth SID of i-th instance, and the $p(\boldsymbol{l}|\boldsymbol{x}_i; \boldsymbol{\theta})$ is the predicted SID. The second term is a regularizer, where N is the set of the indices of the neighboring labels, and $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}_n$ represent the rows in $\boldsymbol{\theta}$ corresponding to the m-th and n-th labels, respectively. Suppose each label is a D-dimension vector, then the labels l_i and l_j are neighbors if $|l_i^1 - l_j^1| \leq \delta_1 \wedge |l_i^2 - l_j^2| \leq \delta_2 \wedge ... \wedge |l_i^D - l_j^D| \leq \delta_D$, where l_i^k and l_j^k are k-th element in l_i and l_j, respectively. $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_D)$ is a D-dimension parameter vector that controls the distance among neighbors. The regularization term guarantees that similar labels have similar supervision intensities, which makes the SID smooth. λ_1 is a balance factor.

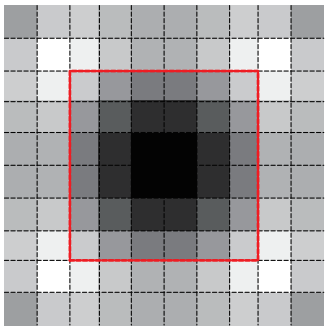


Figure 3: Supervision information expansion through SID. Labels inside red box are unseen labels. For each pixel, higher intensity (brighter) means strong supervision information of the corresponding label.

[1]The label considered in this paper might be multivariate, such as the 2-dimensional head pose label used in the later experiments

[2]In this paper, we only do experiments on databases which labels are numbers.

(a) Pointing'04 database (b) BJUT-3D database (c) MORPH database

Figure 4: Examples in the Pointing'04 database, BJUT-3D database and MORPH database.

Substituting Eq. (5) to Eq. (6), we get the target function

$$T(\boldsymbol{\theta}) = \sum_i \ln \Gamma_i - \sum_i \sum_j (y_i^j \sum_r \theta_{jr} x_i^r) + \sum_i \sum_j \frac{1}{\Gamma_i} \exp(\sum_r \theta_{jr} x_i^r)(\sum_r \theta_{jr} x_i^r - \ln \Gamma_i - \ln y_i^j) + \frac{1}{2}\lambda_1 \sum_{(m,n)\in N} ||\boldsymbol{\theta}_m - \boldsymbol{\theta}_n||_2^2. \quad (7)$$

The minimization of $T(\boldsymbol{\theta})$ can be solved by the limited-memory quasi-Newton method L-BFGS [Liu and Nocedal, 1989]. After obtaining the optimal parameter $\boldsymbol{\theta}^*$, we can get the predicted SID by $p(\boldsymbol{l}|\boldsymbol{x}'; \boldsymbol{\theta}^*)$ given instance $\boldsymbol{x}'$. The label corresponding to the maximum supervision intensity is regarded the final label prediction for $\boldsymbol{x}'$.

3 Experiments

In this section, we will report experiments on two computer vision tasks, i.e., head pose estimation and facial age estimation. For each task, we test not only the performance of the proposed ordinal zero-shot learning algorithm, but also its robustness against the increase of unseen labels in different patterns.

3.1 Datasets

For head pose estimation, the datasets used in this experiment are the Pointing'04 database [Gourier *et al.*, 2004] and the BJUT-3D Chinese Face database [Baocai *et al.*, 2009]. For age estimation, the dataset is the MORPH database [Ricanek and Tesafaye, 2006]. Three examples in these databases are shown in Figure 4.

The Pointing'04 database has 13 yaw angles $\{\pm 90°, \pm 75°, \pm 60°, \pm 45°, \pm 30°, \pm 15°, 0°\}$ and 9 pitch angles $\{\pm 90°, \pm 60°, \pm 30°, \pm 15°, 0°\}$. A pose is represented by the combination of a yaw angle and a pitch angle. Specially, when the pitch angle is $\pm 90°$, the yaw angle is always $0°$. Thus, there are in total $13 \times 7 + 2 = 93$ poses involved in the dataset. The images are taken from 15 different human subjects in two different time and there are $93 \times 15 \times 2 = 2790$ images. For each image, the bounding box of the face is provided in the database.

The BJUT-3D Chinese Face database is a three dimension face database including 500 Chinese persons. There are 250 females and 250 males in this database. Everyone has a high-resolution 3D face data which is acquired by the CyberWare 3D scanner. We render 65 different poses from each 3D face which has 9 yaw angles $\{\pm 60°, \pm 45°, \pm 30°, \pm 15°, 0°\}$ and 9 pitch angles $\{\pm 40°, \pm 30°, \pm 20°, \pm 10°, 0°\}$. Specially, when the pitch angle is $\pm 40°$, the yaw angle is always $0°$ because the 3D faces only include the frontal surface so that too large angles will fail the rendering process. Thus there are $500 \times 65 = 32500$ images in total.

The MORPH database is one of the largest publicly available aging face databases. It contains 55,132 face images from more than 13,000 subjects and the ages of the face images range from 16 to 77. The MORPH database is very imbalanced with regard to the number of instances for each age. Generally speaking, the face images from the ages older than 45 are relatively rare in this database. In order to eliminate the influence of class imbalance, we only use the instances with age labels between 16 to 45, which results in a total of 47,579 instances.

	Pointing'04 database	BJUT-3D database
SIDL	$\mathbf{6.83}°$	$\mathbf{1.80}°$
Kernel PLS	$7.73°$	$2.08°$
Linear PLS	$14.25°$	$4.57°$
Kernel SVR	$11.15°$	$2.17°$
Linear SVR	$13.96°$	$2.32°$

Table 1: Ordinal zero-shot learning results on Pointing'04 database and BJUT-3D database. Lower result is better.

	MAE
SIDL	**3.88**
Kernel PLS	4.12
OHRank	5.82
AAS	4.54
WAS	8.49
Kernel SVR	8.14
CART	5.26

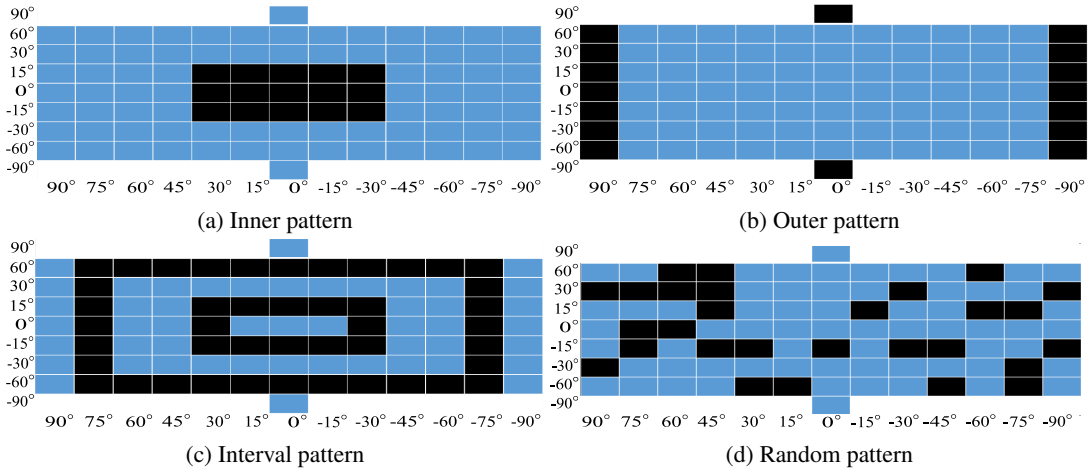
Table 2: Ordinal zero-shot learning results on MORPH database. Lower result is better.

3.2 Zero-Shot Learning Experiment

On each dataset, we randomly choose 2/3 labels as seen labels in the training phase. Then, we test on the data whose labels are the remaining 1/3 labels.

For head pose estimation, we resize each image to 64×64 and the features are extracted by a three-level pyramid HOG of cell sizes: 8×8, 16×16 and 32×32. The SID is generated according to Eq. (2) and $\boldsymbol{\Sigma} = \begin{bmatrix} \tau_1^2 & 0 \\ 0 & \tau_2^2 \end{bmatrix}$, where $\tau_1 = \tau_2 = 30$ for the Pointing'04 dataset, which are 2 times of the minimum angle interval ($15°$). $\boldsymbol{\delta}$ is set to $(15, 15)$. These hyper-parameters are obtained by cross validation. Similarly, we set $\tau_1 = 20, \tau_2 = 30$ on the BJUT-3D dataset and $\boldsymbol{\delta}$ is $(10, 15)$. In this experiment, we compare several regression methods including Linear/Kernel PLS [Haj *et al.*, 2012], Linear/Kernel SVR [Guo *et al.*, 2008]. As to the parameter configurations, on the Pointing'04 dataset, Kernel PLS uses the RBF kernel with the width of 3; Kernel SVR is implemented by by LIBSVM [Chang and Lin, 2011] using the RBF kernel with the parameter "gamma" of 0.01; On the BJUT-3D dataset, Kernel PLS uses the RBF kernel with the kernel width of 5; for Kernel SVR, we use RBF kernel with the "gamma" of 0.1.

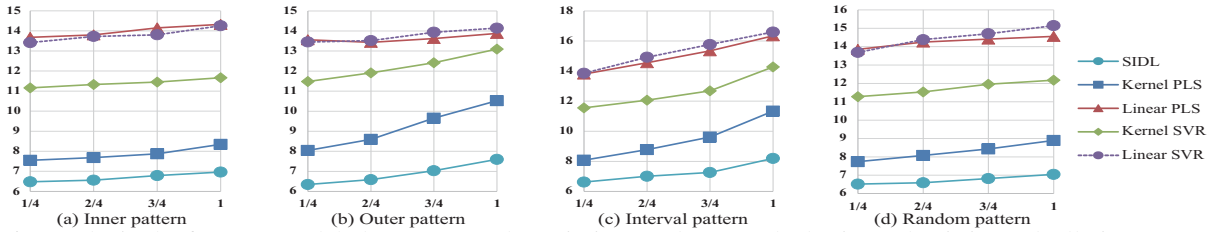
For age estimation, the features extracted from the face images are the Biologically Inspired Features (BIF) [Guo *et al.*,



(a) Inner pattern (b) Outer pattern

(c) Interval pattern (d) Random pattern

Figure 5: Four unseen label patterns for the Pointing'04 dataset. Blue parts represent seen labels, black parts represent candidate unseen labels. Unseen labels will be selected from these candidate unseen labels. Better view in color.



(a) Inner pattern (b) Outer pattern (c) Interval pattern (d) Random pattern

Figure 6: Results in the four unseen label patterns on the Pointing'04 dataset. The horizontal axis is gradually increase number of unseen labels, vertical axis is MAE. Better view in color.

2009] and their dimensionality is reduced to 200 by Marginal Fisher Analysis (MFA) [Yan *et al.*, 2007]. The SID of each face image is initialized using Eq. (2) with its mean at the chronological age and $\mathbf{\Sigma}$ degenerates to variance, which is set to 7^2. $\boldsymbol{\delta}$ in this dataset is a 1-dimensional vector and is set to 1. Several existing algorithms are compared as the baseline methods, which include AAS [Lanitis *et al.*, 2004], WAS [Lanitis *et al.*, 2002], OHRank [Chang *et al.*, 2011], and Kernel PLS [Guo and Mu, 2011]. Two conventional general-purpose regression methods are also compared, i.e., Kernel SVR and CART (Classification and Regression Tree). For AAS, the error threshold in the appearance cluster training step is set to 3. For OHRank, the absolute cost function and the RBF kernel are used. Kernel PLS uses the RBF kernel with the width of 1. Kernel SVR uses the RBF kernel with the "gamma" of 15. CART is implemented as the "regression" type and set to the default values of the MATLAB implementation ("classregtree" class in MATLAB).

The performance of head pose estimation and age estimation is measured by the commonly used mean absolute error (MAE). For head pose estimation, the MAE of yaw and pitch is calculated by the Euclidean distance between the predicted (pitch,yaw) pair and the ground truth (pitch,yaw) pair. For age estimation, the MAE is the average absolute difference between the estimated age and the chronological age. The results of the two tasks are shown in Table 1 and Table 2.

As can be seen, SIDL performs significantly better than other baseline methods. For example, compared to the second best result, the MAE of SIDL decreases by 11.64%, 13.46% and 5.83% on three datasets, respectively. The advantage of SIDL on age estimation is not as prominent as that on head pose estimation. This is because that in head pose estimation, there are more labels with higher dimensionality so that supervision information can be expanded more effectively among the labels.

3.3 Robustness Experiment

In previous experiments, SIDL appears robust against missing class labels since it expands supervision information to the unseen labels. To further demonstrate this, we design an experiment that gradually increases the number of unseen labels. In order to observe the performance variation of a particular algorithm, the test set should keep the same when the unseen labels in the training set changes. So we first split the dataset into a training set and a test set. Then, while keeping the test set unchanged, the instances associated to more and more labels in the training set are gradually removed.

For each dataset, first we generate a candidate unseen label set and then we randomly select labels as unseen labels from the candidate unseen label set. As discussed in Section 2.1, we generate candidate label set according to the four unseen label patterns shown in Figure 2. In detail, Figure 5 shows the candidate unseen label sets for different patterns on the Pointing'04 dataset (those for the BJUT-3D dataset and MORPH dataset are similar). Black parts represent candidate unseen labels and blue parts represent seen labels. Then, for head pose estimation, in each pattern we choose the unseen labels 4 times and the proportions of the unseen labels to the total candidate unseen labels gradually increase as $1/4$, $2/4$, $3/4$ and 1; for age estimation, we choose the unseen labels 3 times

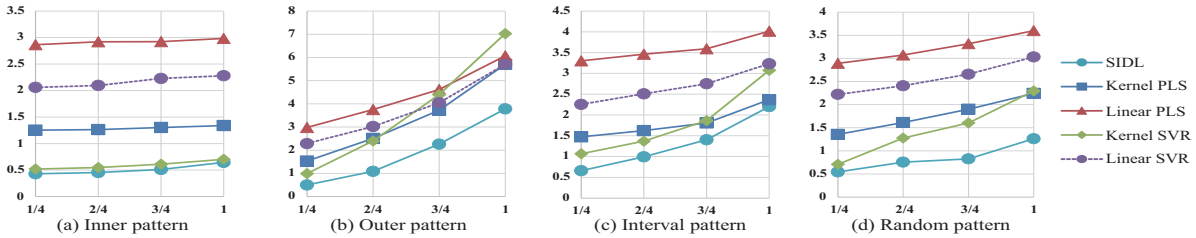


Figure 7: Results in the four unseen label patterns on the BJUT-3D dataset. The horizontal axis is gradually increase number of unseen labels, vertical axis is MAE. Better view in color.

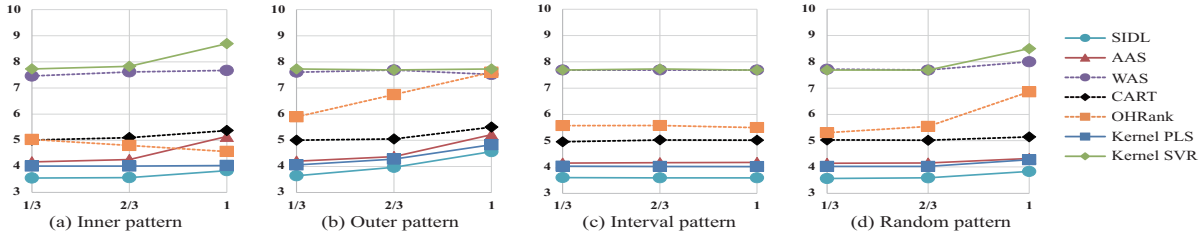


Figure 8: Results in the four unseen label patterns on the MORPH dataset. The horizontal axis is gradually increase number of unseen labels, vertical axis is MAE. Better view in color.

and the proportions are $1/3$, $2/3$ and 1, respectively. Note that the unseen labels are randomly chosen without replacement, i.e., the latter selection will include the previous selection.

Each algorithm undergoes a five-fold cross validation on each unseen label pattern. The performance with the increase of unseen labels on the three datasets is shown in Figure 6, Figure 7 and Figure 8, respectively. Generally speaking, SIDL performs the best in all cases compared with the baseline methods. More importantly, in most cases, the performance of SIDL deteriorates most slowly with the increase of unseen labels. This indicates that SIDL deals with ordinal zero-shot learning well in not only the value of MAE, but also the robustness against more and more unseen labels. Note that when starting from the lowest MAE, it is much harder for SIDL to remain the slowest performance deteriorating rate than those baseline methods starting from higher MAE. For example, on the BJUT-3D dataset, the second best algorithm when 1/4 unseen labels are selected, Kernel SVR, deteriorates rapidly with the increase of unseen labels. On the other hand, although some algorithms starting from a very high MAE, such as Linear PLS, and deteriorate slowly, it does not make much sense since their performance is significantly worse than SIDL consistently. Note that exceptions might happen only in the inner and interval patterns on the BJUT-3D and the MORPH datasets. For example, SIDL performs similarly with Kernel SVR in the inner pattern on the BJUT-3D dataset. Its performance deteriorates faster than Kernel PLS in the interval pattern on the BJUT-3D dataset and in the inner pattern on the MORPH dataset. The reason might be that in both the inner and interval patterns, the unseen labels are surrounded by the seen labels. So they are relatively easier cases compared with other patterns. Thus, general-purpose regression methods like SVR or PLS may also achieve good performance. Even though, SIDL still performs better than their best achievements.

Moreover, it can be observed that SIDL exhibits different performance trends in different patterns. For example, on all datasets, the trends in the inner, interval and random patterns are usually smoother than those in the outer pattern. This is because that in the former three patterns, the unseen labels are often surrounded by other seen labels, and thus the label correlation can be better utilized to help with the learning of unseen labels.

4 Conclusion

This paper is motivated by the missing ordinal labels in the training set. Ordinal labels are usually closely related to each other, which we can use to expand the supervision information in zero-shot learning. Toward this, we propose to use Supervision Intensity Distribution (SID) in zero-shot learning for ordinal classification without side information. SID contains the supervision intensity of each label to a certain instance. Learning from SID is mainly implemented by minimizing the Jeffrey's divergence between the ground truth SID and predicted SID. By conducting experiments on head pose and facial age estimation, we demonstrate that our method performs significantly better than compared regression methods, and is more robust against the increase of unseen labels.

Acknowledgments

This research was supported by National Science Foundation of China (61622203, 61232007), Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140022), Collaborative Innovation Center of Novel Software Technology and Industrialization, and Collaborative Innovation Center of Wireless Communications Technology.

References

[Akata *et al.*, 2013] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, Portland, Oregon, USA, June 2013.

[Baocai *et al.*, 2009] Yin Baocai, Sun Yanfeng, Wang Chengzhang, and Ge Yun. Bjut-3d large scale 3d face database and information processing. *Journal of Computer Research and Development*, 6:020, 2009.

[Berg *et al.*, 2010] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *in Proc. European Conf. Computer Vision*, pages 663–676, Crete, Greece, 2010. Springer.

[Branson *et al.*, 2010] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *in Proc. European Conf. Computer Vision*, pages 438–451, Crete, Greece, 2010. Springer.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Chang *et al.*, 2011] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, pages 585–592, Colorado Springs, CO, 2011.

[Elhoseiny *et al.*, 2013] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *in Proc. Intl Conf. Computer Vision*, 2013.

[Farhadi *et al.*, 2009] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, pages 1778–1785, Miami, Florida, USA, 2009. IEEE.

[Geng and Xia, 2014] Xin Geng and Yu Xia. Head pose estimation based on multivariate label distribution. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, pages 1837–1842, Columbus, OH, USA, 2014.

[Geng *et al.*, 2013] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

[Geng, 2016] Xin Geng. Label distribution learning. *IEEE Trans. on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[Gourier *et al.*, 2004] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Proc. Pointing 2004, ICPR, Int'l Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK, 2004.

[Guo and Mu, 2011] Guodong Guo and Guowang Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, pages 657–664, Colorado Springs, CO, 2011.

[Guo *et al.*, 2008] Guodong Guo, Yun Fu, Charles R. Dyer, and Thomas S. Huang. Head pose estimation: Classification or regression? In *in Proc. Int'l Conf. Pattern Recognition*, pages 1–4, Tampa, FL, 2008.

[Guo *et al.*, 2009] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S. Huang. Human age estimation using bio-inspired features. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, pages 112–119, Miami, FL, 2009.

[Haj *et al.*, 2012] Murad Al Haj, Jordi Gonzàlez, and Larry S. Davis. On partial least squares in head pose estimation: How to simultaneously deal with misalignment. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, pages 2602–2609, Providence, RI, 2012.

[Lampert *et al.*, 2014] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[Lanitis *et al.*, 2002] A. Lanitis, C. J. Taylor, and T.F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[Lanitis *et al.*, 2004] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004.

[Liu and Nocedal, 1989] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(3):503–528, 1989.

[Palatucci *et al.*, 2009] Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *in Proc. Advances in Neural Information Processing Systems*, pages 1410–1418, Vancouver,Canada, 2009. Curran Associates, Inc.

[Parikh and Grauman, 2011] Devi Parikh and Kristen Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, pages 1681–1688, Colorado Springs, CO, 2011. IEEE.

[Ricanek and Tesafaye, 2006] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *in Proc. Int'l Conf. Automatic Face and Gesture Recognition*, pages 341–345, Southampton, UK, 2006.

[Rohrbach *et al.*, 2010] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where– and why? semantic relatedness for knowledge transfer. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, pages 910–917, San Francisco, California, USA, 2010. IEEE.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *in Proc. Advances in Neural Information Processing Systems*, pages 935–943, South Lake Tahoe, Nevada, US, 2013.

[Yan *et al.*, 2007] Shuicheng Yan, Dong Xu andBenyu Zhang, HongJiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

[Zhang and Saligrama, 2016] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *in Proc. IEEE Intl Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada,USA, June 2016.