Leveraging Group Classification with Descending Soft Labeling for Deep Imbalanced Regression

Ruizhi Pu[1*], **Gezheng Xu**[1], **Ruiyi Fang**[1], **Bing-Kun Bao**[2], **Charles Ling**[1†], **Boyu Wang**[1†]

[1] Department of Computer Sceince, Western University
[2] School of Computer Science, Nanjing University of Posts and Telecommunications

Abstract

Deep imbalanced regression (DIR), where the target values have a highly skewed distribution and are also continuous, is an intriguing yet under-explored problem in machine learning. While recent works have already shown that incorporating various classification-based regularizers can produce enhanced outcomes, the role of classification remains elusive in DIR. Moreover, such regularizers (e.g., contrastive penalties) merely focus on learning discriminative features of data, which inevitably results in ignorance of either continuity or similarity across the data. To address these issues, we first bridge the connection between the objectives of DIR and classification from a Bayesian perspective. Consequently, this motivates us to decompose the objective of DIR into a combination of classification and regression tasks, which naturally guides us toward a divide-and-conquer manner to solve the DIR problem. Specifically, by aggregating the data at nearby labels into the same groups, we introduce an ordinal group-aware contrastive learning loss along with a multi-experts regressor to tackle the different groups of data thereby maintaining the data continuity. Meanwhile, considering the similarity between the groups, we also propose a symmetric descending soft labeling strategy to exploit the intrinsic similarity across the data, which allows classification to facilitate regression more effectively. Extensive experiments on real-world datasets also validate the effectiveness of our method.

Appendix — https://github.com/RuizhiPu-CS/Group-DIR

Introduction

Data imbalance exists ubiquitously in real-world scenarios, posing significant challenges to machine learning tasks as certain labels may be less observed than others or even missed during training. Although the imbalanced problem has been extensively studied in the field of classification (He and Garcia 2009), how to tackle deep imbalanced regression (DIR) is still under-explored.

Due to the continuity of the label space and the dependence of data across nearby targets (Yang et al. 2021), previous solutions in DIR primarily focused on estimating accurate imbalanced label density, such as label distribution smoothing (LDS), feature distribution smoothing (FDS) (Yang et al. 2021) and re-weighting (Cui et al. 2019; Branco, Torgo, and Ribeiro 2017). Meanwhile, (Ren et al. 2022) proposed a balanced Mean Square Error (B-MSE) loss to accommodate the imbalanced distribution in the label space. Recent works incorporated classification regularizers with the Mean Square Error (MSE) in DIR, such as contrastive regularization (Zha et al. 2023a; Keramati, Meng, and Evans 2023), entropy regularization (Zhang et al. 2023), and feature ranking regularization (Gong, Mori, and Tung 2022), which have achieved significant performance improvements. Moreover, (Pintea et al. 2023) formalized the relation between the balanced and imbalanced regression and empirically investigated how classification can help regression.

Although incorporating classification regularizers in DIR has already achieved enhanced output, the relationship between the objectives of classification and DIR remains elusive. In the meantime, these classification regularizers would also force the model to focus more on the discriminative fea-

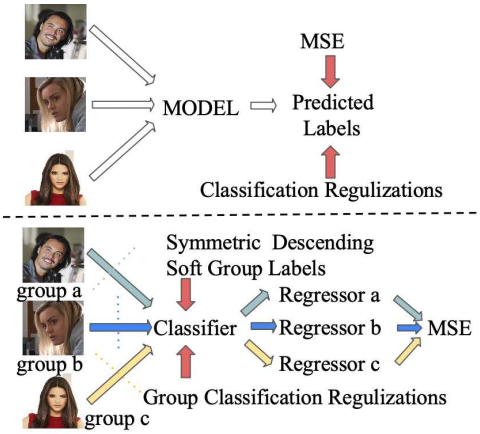

Figure 1: Comparison between previous works and ours. **Upper)** Previous methods directly incorporated classification regulations (Gong, Mori, and Tung 2022; Zhang et al. 2023). **Bottom)** We propose a descending soft labeling to leverage the classification to help DIR (Different colors denote different groups of data).

*Email: rpu2@uwo.ca
†Corresponding author, email: bwang@csd.uwo.ca

ture which is inappropriate for regression tasks. For example, for a facial-image-based age regression task, the images corresponding to nearby labels exhibit both continuity and similarity. A photo of a 40-year-old person should resemble those of both 35-year-olds and 45-year-olds, and also reflect an intermediate stage in age-related features. However, such property (data similarity) has been always ignored in existing classification-based methods.

In this paper, to investigate the connections between the classification and DIR, we revisit the objective of DIR from a Bayesian perspective. We show that the objective of DIR can be decomposed into the combination of both group classification and sample regression within each group. Inspired by this finding, we can explicitly leverage the classification to help DIR in a divide-and-conquer manner.

Specifically, considering that data with nearby labels would naturally be similar (Yang et al. 2021; Pintea et al. 2023) in DIR, we aggregate the data of close labels as the same groups. Hereby, we divide the whole dataset into continuous but disjoint groups and convert the DIR into a classification problem. In the meantime, these divided groups can not only preserve the ordinal information as their original labels but also provide us with a feasible way to explore the connection between the group classification and DIR.

Subsequently, since the decomposition of the DIR objective would also split the imbalance into both objectives of group classification and regression, inspired by (Liu et al. 2021) that feature representations learned by self-supervised learning can exhibit imbalance-robust, we introduce an ordinal group contrastive learning to learn an ordinal high-quality feature representation to build a solid foundation for both classification and regression tasks. Afterward, we make the group prediction for each learned representation on a classifier (Divide). With this group estimation, we employ a multi-experts regressor to regress the representation on its corresponding predicted group (Conquer). The difference between our proposed method and the previous works can be found in Fig.1, where the previous works handle all data simultaneously while our work first divides the data into different groups and then conquers them with each expert regressor given their corresponding groups.

However, empirical observation shows that it is difficult to make an accurate group estimation under standard classification loss such as cross-entropy (CE) loss. For example, in Fig.2, the data samples from group 1 to 5 (minorities) are rarely correctly predicted. Instead, most of the data samples are over-estimated into groups 6,8,12, and 14 (majorities). The primary cause of this inaccurate prediction is the data dependence of nearby groups (images of close groups). (Yang et al. 2021) as each group also exhibits different levels of similarity between each other. As shown in Fig.2, groups with minority data samples would be easily misclassified into their neighboring groups with majority data samples.

As a result, these imprecise group predictions would misguide the data samples into the incorrect expert regressors and result in performance degradation. To tackle this problem, we propose a symmetric descending soft labeling strategy that leverages the intrinsic label similarity of the data for the group prediction. Since the labels can not only present

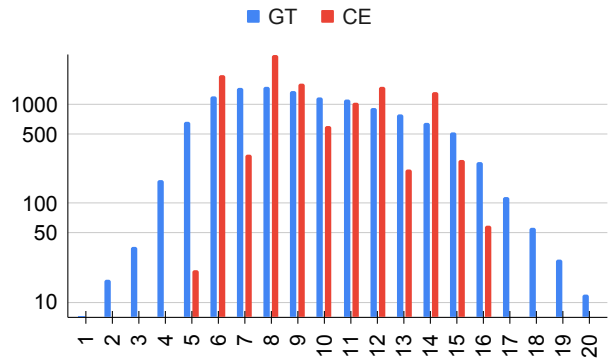

Figure 2: Comparison between the (Logarithm scale) Ground Truth (GT) label and estimated label based on CE. X: group id Y: number of samples.

the discrepancy information but also reflect the relative similarity between the data in DIR, we encode the group label into the soft labels which descend symmetrically from their group label until the end of the groups to capture the similarities between the groups.

In a nutshell, by incorporating the classification with the symmetric descending soft labeling into DIR, we provide a novel framework to address the DIR in a divide-and-conquer manner. More importantly, we also conduct comprehensive experiments with various real-world datasets, demonstrating the effectiveness of our method.

In summary, we conclude our contributions as follows:

- We revisit the objective of DIR from a Bayesian perspective, which motivates us to address the DIR in a divide-and-conquer manner.
- We incorporate an ordinal group-aware contrastive learning to learn a high-quality feature representation to provide a solid foundation for both classification and regression tasks in our decomposed objective.
- We introduce a multi-experts regressor to handle different groups of data with different expert regressors and we propose a symmetric descending soft labeling strategy to capture the similarity across the data in DIR.

Motivation

Preliminary

We study the DIR problem in this paper. In DIR, we assume that we have a training set $\{x_i, y_i\}_{i=1}^N$ with size N, where $x_i \in \mathbb{R}^d$ is the input with dimension d and $y_i \in \mathbb{R}$ is the label. Meanwhile, the distribution of this training set p_{tr} is always highly skewed. The objective of DIR is to learn a model from this highly skewed training set to generalize well on an unseen test set with the balanced distribution p_{bal}. In this paper, we aim to learn a feature extractor f with parameter $\mathbf{w}_f$, a classifier h with parameter $\mathbf{w}_h$ and a set of regressors $\varphi = [\varphi_0, \ldots, \varphi_{|G|-1}]$ with parameter $\mathbf{w}_\varphi = [\mathbf{w}_{\varphi_0}, \ldots, \mathbf{w}_{\varphi_{|G|-1}}]$ simultaneously, the parameters of the model consists of $\theta = \{\mathbf{w}_f, \mathbf{w}_h, \mathbf{w}_\varphi\}$.

Motivation

We first revisit our goal from a Bayesian perspective. In DIR, our goal is to learn a model with parameter θ via the MSE loss to model the imbalanced training distribution $p_{tr}(y|x)$ and generalize well on the unseen balanced test distribution $p_{bal}(y|x)$. Since directly adopting MSE loss in DIR is in fact to model the $p(y|x)$ for an underlying Gaussian distribution (Ren et al. 2022), a model learned from an imbalanced set would consequently underestimate rare labels, limiting its ability to generalize to an unseen balanced set (Groups are mapped from labels, e.g., for a mapping $g = \left\lfloor \frac{y}{|G|} \right\rfloor$).

Therefore, we can review the conditional distribution of training data $p_{tr}(y|x)$ as follows:

Lemma 1 (Group-aware Bayesian Distribution Modeling for DIR)**.** *The conditional distribution of $p_{tr}(y|x)$ in the training of DIR can be decomposed into a combination of both classification and regression tasks summing over distinct groups:*

$$
\begin{aligned}
p_{tr}(y|x) &= \frac{p_{tr}(x,y)}{p_{tr}(x)} = \frac{\sum_{g\in G} p_{tr}(x,y,g)}{p_{tr}(x)} \\
&= \frac{\sum_{g\in G} p_{tr}(g|x)p_{tr}(x)p_{tr}(y|x,g)}{p_{tr}(x)} \\
&= \sum_{g\in G} p_{tr}(g|x)p_{tr}(y|x,g)
\end{aligned}
$$

where G is the set of groups, and $|G|$ is the number of groups, and we abbreviate g as the group label.

We take a step forward by taking negative logarithm at both sides [1], we can obtain the learning objective of DIR in the form of loss as:

$$
\begin{aligned}
-\log p_{tr}(y|x) &\leq \sum_{g\in G} -\log(p_{tr}(g|x)p_{tr}(y|x,g)) \\
&= \sum_{g\in G} \underbrace{-\log p_{tr}(g|x)}_{groups\ classification} \underbrace{-\log p_{tr}(y|x,g)}_{labels\ regression}
\end{aligned} \tag{1}
$$

Remark: The learning objective of DIR can be decomposed into two perspectives, 1) the objective of imbalance group classification to predict the group label, 2) the objective of imbalance regression to regress the data labels, showcasing that we can solve the DIR in a divide-and-conquer manner. Empirical results from Fig.3 also validate the effectiveness of objective decomposition in Lemma 1 for addressing DIR.

As we can observe from Fig.3[2], if we train a vanilla model with MSE loss only (regression-only), the training MSE loss curve and the validation MSE curve converges with different scales and the convergence speed diverges a lot, demonstrating that the training from the imbalanced set would result in unsatisfying results on the balanced validation set due to the imbalance in the training set.

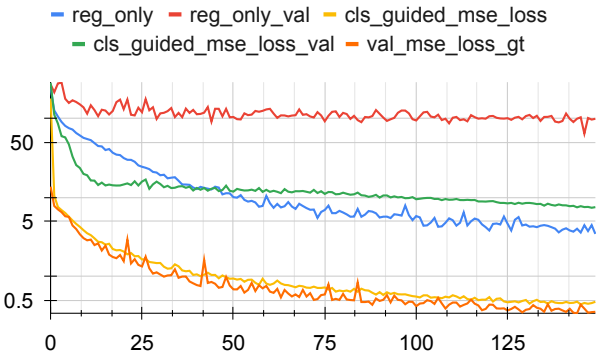

Figure 3: MSE results between model trained with MSE and model trained with the decomposition loss from Lem. 1 (20 groups) on imbalanced Train & balanced Validation set (AgeDB-DIR). Row: Epoch, Column: MSE (Note: column is in logarithmic scale for an easier observation).

Instead, when we substitute the MSE loss with classification loss $p(g|x)$ (e.g. CE) and the classification-guided MSE loss $p(y|x,g)$ as Lemma 1, the classification-guided MSE loss exhibits a sharp converge compared with the vanilla model. Moreover, the validation MSE of the classification-guided regression also converges more sharply compared with that of the vanilla model and not even at the same scale, demonstrating that the representation learned from classification can help to address the DIR and showcasing the effectiveness of alleviating the negative impact of imbalance in DIR with our classification guided regression. This motivates us to perform the divide-and-conquer by first estimating the groups $p(g|x)$ and then guiding the group-corresponding-regressors to regress the labels of the data samples $p(y|x,g)$.

Hereby, we connect the classification with the objective of the DIR (to model $p_{tr}(y|x)$ from $p_{tr}(y|x;\theta)$). Furthermore, the above lemma also demonstrates that the objective of DIR can be upper-bounded by both classification and regression. By minimizing the empirical risk of the classification of the groups (to model $p_{tr}(g|x)$ with $p_{tr}(g|x;\mathbf{w}_f,\mathbf{w}_h)$) and guiding the predictions of groups to minimize the empirical risk of the regression (to model $p_{tr}(y|x,g)$ with $p_{tr}(y|x,g;\mathbf{w}_f,\mathbf{w}_\varphi)$) simultaneously, we can properly address the DIR from a Bayesian perspective. More importantly, this motivates us to solve the DIR problem in a divide-and-conquer manner as we can leverage the group classification $p(g|x)$ to guide the learning of regression $p(y|x,g)$.

Methodology

In this section, we introduce ordinal group-aware constrastive learning to learn a high-quality feature representation which is beneficial for both classification and regression. Then, we leverage a multi-experts regressor to conduct regression under the guidance of the group predictions to

[1] the inequality comes from $\log(a+b) \geq \log a + \log b$ for $0 < a < 1$ and $0 < b < 1$ since the elements of logarithm in above are probabilities which less than 1, the inequality holds.

[2] The _val denotes the validation performance, _train denotes the training performance, cls_guided denotes the data sample is regressed on the predicted regressor and gt_guided denotes the data sample is regressed on the its true regressor.

fully exploit the benefits from classification to help regression in a divide-and-conquer manner. Furthermore, we propose a symmetric descending soft labeling strategy to capture the data similarity across groups.

Ordinal Group-aware Contrastive Learning

As in DIR, label space is not only continuous but also ordinal. Consequently, we introduce an ordinal group-aware contrastive learning to learn a high-quality feature representation. Meanwhile, this high-quality representation can also act as a solid foundation for both classification $p(g|x)$ and the regression $p(y|x,g)$ tasks as described in our objective decomposition from Lemma 1.

Inspired by (Zha et al. 2023b; Xiao et al. 2023), we introduce ordinal contrastive learning in a group-aware manner. Since data samples at nearby labels would have similar features (e.g., facial samples from 30 to 40 would be similar to each other), we cluster the data samples with their corresponding groups. Different from (Zha et al. 2023b), we concentrate on investigating relationships between these groups to help to learn a high-quality feature representation.

As these distinct groups would preserve the ordinal as their original labels (e.g., the label of arbitrary sample in group 0 would always be smaller than the label of arbitrary sample in group 1), we focus on constructing an ordinal group-aware contrastive learning framework where the learned feature representations can also preserve this ordinal characteristics between the groups. To achieve this goal, for an anchor group label i and another arbitrary group label j, we push away other samples whose group label distances are more distant than i and j. If two samples are in the same group, we pull them together at the feature space. In this way, data samples with different distances in group labels would be pushed away in various degrees, as the closer groups would be pushed less than the distant groups.

Hereby, we formulate the ordinal group-aware contrastive loss as the following:

$$\mathcal{L}_{grc}(\mathbf{w}_f) = - \frac{1}{B(B-1)} \sum_{i=1}^{B} \sum_{\substack{j=1, \\ j \neq i}}^{B} \log \frac{s(z_i, z_j)}{\sum_{k=1}^{B} \mathbf{1}_{[\phi(i,j,k)]} s(z_i, z_k)} \tag{2}$$

where for the index i, j, k of three arbitrary data samples in a batch, z is the feature representation, $s(i,j)$ is the abbreviate of $exp(sim(z_i, z_j)/t)$ and $sim(\cdot)$ denotes the similarity function (e.g., cosine similarity), $exp(\cdot)$ is the exponential function, $\phi(i,j,k) \triangleq \{k \neq i, d(g_i, g_k) \geq d(g_i, g_j)\}$ is the condition of the zero-one indicator $\mathbf{1}$ (return 1 where ϕ satisfies and 0 vice verse), g denotes the group label of the data sample, t is the temperature hyper-parameter, B is the batch size, and $d(\cdot)$ denotes the distance function (e.g., L1 distance). By comparing the relative distance of group labels between two arbitrary samples, we can achieve the group ordinal as that of the labels in the feature space.

Classification Guided Multi-experts Regression : Modeling $p(y|x,g)$

With the acquired contrastive representations, we introduce a multi-experts regressor to tackle each group of data in a divide-and-conquer manner. In the training phase, given the ground truth group label of each data sample, we perform regression on its corresponding expert regressor. In the testing phase, each data sample is first classified into a group. Since each predicted group corresponds to an expert regressor, we conduct regression on the predicted expert regressor during the testing phase.

Therefore, we formulate the multi-expert regression MSE loss as follows:

$$\mathcal{L}_{mse}(\mathbf{w}_f, \mathbf{w}_\varphi) = \sum_{g=0, y \in [g]}^{|G|-1} (y_{\varphi_g} - \hat{y}_{\varphi_g})^2 \tag{3}$$

where $y \in [g]$ denotes the label y belongs to group g. We have $\hat{y}_{\varphi_g} = \mathbf{w}_{\varphi_g}(z_{\varphi_g})$ for the group of data samples whose ground truth group labels are g and their learned representations z are then forwarding to their corresponding regressors φ_g to obtain the prediction $\hat{y}_{\varphi_g}$. We abbreviate the ground truth target labels as y_{φ_g}. Moreover, LDS can also be utilized to further tackle the intra-group imbalance. Since the final MSE is calculated on each data sample and each data sample corresponds to each group, we accumulate the MSE loss over all groups.

Symmetric Descending Soft Labeling for Group Classification : Modeling $p(g|x)$

However, since the nearby label data would exhibit data dependence (Yang et al. 2021) in DIR, in our framework, the nearby group data would also exhibit data dependence. Consequently, the data dependence of nearby groups and inherent group imbalance would hinder us from making accurate group estimations for regression.

When we directly utilize the standard CE loss to estimate the group label of the data, as can be observed from Fig.2, the predictions mostly fall into the groups with the majority of samples. Meanwhile, when we adopt logits adjustment (LA) (Menon et al. 2021), which is one of the most effective imbalance classification solutions, to predict the group labels, another empirical observation arises in Fig.4 that this method over-estimate the groups with minority samples.

Therefore, empirical results in Fig.2 and Fig.4 have shown that classification loss such as CE and LA perform poorly in group prediction of DIR. More importantly, as can be observed from Fig.5, in both CE and LA, the data dependence of the groups would lead the predictions to mainly fall into nearby groups (high absolute difference of misclassification in Fig.5). The reason for this is the classification solutions would focus on the discriminative information (as we stated above) while ignoring the data similarities across the groups.

In standard classification loss such as CE, the ground truth for one group label g is encoded as a vectorized label $l_{gt} = [\ldots, 0, 0, 1, 0, 0, \ldots]$, where 1 is at the position of g-th index, and 0 is at the rest indexes and the CE loss $\mathcal{L}_{ce} = -\log p^g$ is only calculated on the prediction at the index g for the group prediction $p = [p^0, \ldots, p^g, \ldots, p^{|G|-1}]$.

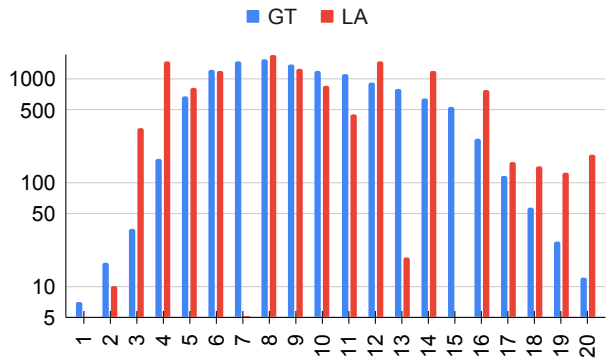

Figure 4: Comparison between the (Logarithm scale) Ground Truth (GT) label and estimated label based on LA. X: group id Y : number of samples.

As a result, only the information on the ground truth index is provided (set to 0) while others are all overlooked (set to 0). However, since the group label is not merely continuous, it's also essential to recognize that labels within DIR encode intrinsic relative similarities between them. Thus, directly adopting classification loss in group prediction is not applicable in DIR.

Therefore, we introduce a symmetric descending soft labeling strategy into the group classification to fully exploit the similarity nature of the groups. To convert a scalar group label into a vectorized label for training, for a group with ground truth label g, we assign the g-th index in the label vector with the highest value of $|G|$ and decrease it symmetrically from the position of the current index until the end. Thus, the soft label of the scalar group label g would be encoded as $l_{soft} = [\ldots, |G|-2\beta, |G|-\beta, |G|, |G|-\beta, |G|-2\beta, \ldots]$, where, $|G|$ is at the index of g in the label vector, β is a hyper-parameter e.g., $\beta = 1$ and it denotes the relative distance between two neighboring labels. We formulate the soft label q_{soft} of a data sample from the ground truth group label as: $q_{soft} = \sigma(l_{soft})$ where σ denotes the SoftMax function, as $\sigma(q_i) = \frac{e^{q_i}}{\sum_{j=1}^{|G|} e^{q_j}}$. Moreover, we briefly show two extreme cases for our soft label, in the case when $g = 0$, the $l_{gt\_soft} = [|G|, |G|-1, |G|-2, \ldots, 1]$, and in the case when $g = |G|$, the $l_{gt\_soft} = [1, \ldots, |G|-2, |G|-1, |G|]$.

The soft label cross-entropy loss for a data sample with group label g (corresponding with the regressor g) in a batch B as the following:

$$\mathcal{L}_{soft}(\mathbf{w}_f, \mathbf{w}_{\varphi_g}) = \sum_{j=1}^{B} \sum_{g=0}^{|G|-1} q_j^g \log p_j^g \tag{4}$$

where p_j^g denotes group prediction and q_j^g is the soft label in q_{soft} of sample j at index g.

By encoding the ground truth labels into soft labels, we can preserve the relative group information of all groups in one single label, providing comprehensive data information for the group classification and also contributing to the regression. Comparison between different classification criteria (Soft Labeling/CE/LA) also shows the effectiveness of our proposed method.

Final Loss

By aggregating the above losses together, the final objective of our proposed method is :

$$\mathcal{L}_{final} = \mathcal{L}_{grc} + \lambda_1 \mathcal{L}_{mse} + \lambda_2 \mathcal{L}_{soft} \tag{5}$$

where λ_1 and λ_2 are hyper-parameters to balance the losses.

Experiments

Datasets

We validate our proposed method with the following real-world dataset which includes both visual tasks and natural language processing tasks:

IMDB-WIKI-DIR is a large-scale real-world human facial dataset constructed by (Rothe, Timofte, and Van Gool 2018) and re-organized for imbalance tasks by (Yang et al. 2021), it contains 235K face images. There are 191.5K imbalance training images, 11K balanced validation images, and 11K balanced test images. The dataset was manually divided given the bin length of 1 year (each bin can be regarded as the target label as in (Yang et al. 2021)).

AgeDB-DIR is another real-world human facial dataset constructed by (Moschoglou et al. 2017) and also re-organized by (Yang et al. 2021). It contains 12.2K image training data, 2.1K image validation data, and 2.1K image test data. The bin length is also 1 year but the minimum age is 0 and the maximum age is 101.

STS-B-DIR is a text similarity score dataset constructed by (Wang et al. 2018) and re-constructed by (Yang et al. 2021). It is collected from news headlines, videos, image captions, and natural language inference data. The dataset is a set of sentence pairs annotated with an average similarity score, and the range of scores varies from 0 to 5. There are 5.2K pairs for the training, 1K balanced pairs for validation, and 1K balanced pairs for test. Each bin length is 0.1.

Implementation Details

Baselines and experiment set up We conducted our experiments with the backbone based on ResNet-50 for AgeDB-DIR & IMDB-WIKI-DIR dataset. For STS-B-DIR, we follow the same standard experiment setting as in (Yang et al. 2021; Ren et al. 2022), we adopted the BiLSTM + GloVe word embeddings and preprocessed them in the experiment. Moreover, we follow the training procedures and hyper-parameters (e.g., temperature t) as (Zha et al. 2023a), but apart from (Zha et al. 2023a) which only used a sub-sample of both datasets (e.g., 32K for IMDB-WIKI-DIR), we stick to the setting of (Yang et al. 2021) and use the full training set with the batch size of 128 for training. Same as (Yang et al. 2021; Branco, Torgo, and Ribeiro 2017), the train data distribution is always highly skewed while the test distribution is balanced. More details can be found in the Appendix.

Result Analysis

AgeDB-DIR : In the dataset AgeDB-DIR, it is obvious that our method outperforms most of other methods in Tab.1.

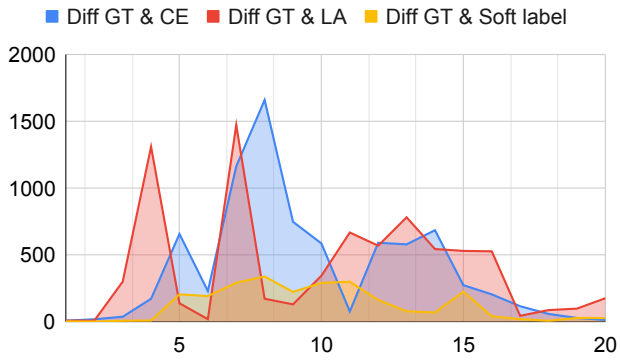

Figure 5: Comparison of the absolute difference (Diff) between group predictions and ground truth in CE, LA, and Ours on AgeDB-DIR. Lower denotes more accurate group predictions. X : group numbers. Y : Absolute value.

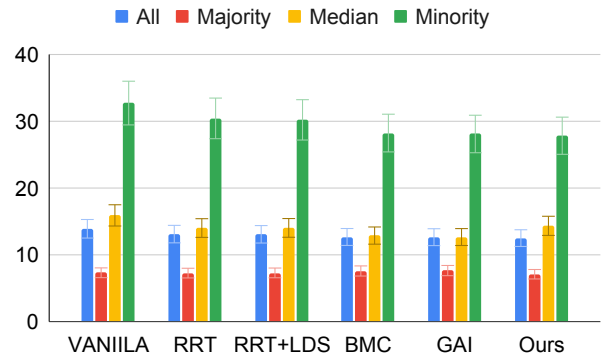

Figure 6: Comparison between various DIR solutions of b-MAE Results in Majority, Median, and Minority on IMDB-WIKI-DIR. Y : b-MAE.

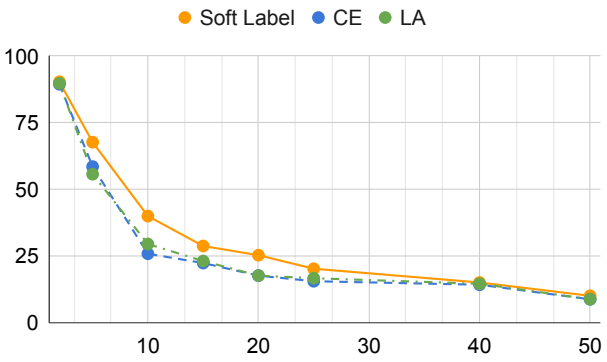

Figure 7: Group prediction accuracy comparison between our Soft labeling/CE/LA on AgeDB-DIR, X: group number Y: group prediction accuracy.

In particular, we show that our method can better deal with the majority and median, without greatly sacrificing the performance of the minorities as in the previous works. Compared with existing works, our work achieved a state-of-art (SOTA) performance with an overall MAE of 6.87. Meanwhile, our work has lower GM, which shows that our results of MAE are averagely smaller than other works, showing the effectiveness of our proposed method.

We show that in Fig.5, we use the absolute difference between the ground truth labels and the estimated labels to identify if our proposed method can help the classification (how accurate the group estimation can be given the ground truth). Our proposed symmetric descending soft labeling significantly outperforms others in group estimation (compared with Fig.2 & 4), that is because the soft labels can help the representations to fully exploit the similarity characteristics of the data from other labels. Consequently, it contributes to a more accurate group estimation than other existing works, resulting in minimizing the $\log p(g|x)$ and the gap Δ at the same time.

Another interesting observation from Fig.5 arises that, directly using the CE and LA would also make the predictions in the tail groups almost fail, that is because the tail groups are always the minorities. Also, our Soft labeling can capture the information from other groups to help minorities in group imbalance. As in Fig.7, the classification performance of the soft labeling is consistently better than that of the CE and LA, such as in 20 and 25 groups, the soft labeling has a 5% more improvement compared to others, which validates that the label similarity is one crucial characteristic in DIR and leverage the group similarities as that of the label similarities can help to take advantages of classification in helping DIR as (Pintea et al. 2023).

IMDB-WIKI-DIR: In the dataset IMDB-WIKI-DIR, which is also the largest DIR real-world dataset, our overall performance in Tab.2 achieved a satisfying result and is better than most of the current solutions. Specifically, we show that our method can have a better performance on the median and the few shot, it shows that our proposed method can exploit more information on the median and the few shots with the soft-label, resulting in an overall performance improvement. Moreover, we show that our result in GM is also better than others both in the Med. and Few., showing a consistent superiority of our work over others and validating that our proposed method can better address the Med. and Few shots in the extreme case of imbalance. As we can observe from Fig.6, the comparison of b-MAE results between the state-of-art DIR solutions and our proposed method also validates the effectiveness of our method, showcasing our method can have a better performance on the balanced sets, especially on the minority samples.

STS-B-DIR: In the dataset STS-B-DIR, it is easy to observe that the under-represented median & minority data samples significantly outperform others in Tab.3, contributing to the overall performance enhancement. Moreover, our work simultaneously improved the performance of median & minority shots in Pearson Correlation compared to others, that is because our soft labeling can help us to preserve the data similarity as that of labels in representation learning, which can enhance the Pearson correlations and consistent

Method \ Shot	MAE↓				GM↓			
	All	Many.	Med.	Few.	All	Many.	Med.	Few.
VANILLA	7.77	6.62	9.55	13.67	5.05	4.23	7.01	10.75
SMOTER	8.16	7.39	8.65	12.28	5.21	4.65	5.69	8.49
SMOGN	8.26	7.64	9.01	12.09	5.36	4.90	6.19	8.44
RRT	7.74	6.98	8.79	11.99	5.00	4.50	5.88	8.63
RRT+LDS	7.72	7.00	8.75	11.62	4.98	4.54	5.71	8.27
FOCAL-R	7.64	6.68	9.22	13.00	4.90	4.26	6.39	9.52
SQINV	7.81	7.16	8.80	11.20	4.99	4.57	5.73	7.77
SQINV + LDS	7.67	6.98	8.86	10.89	4.85	4.39	5.80	7.45
LDS+FDS	7.55	7.01	8.24	10.79	4.72	4.36	5.45	6.79
VAE	7.63	6.58	9.21	13.45	4.86	4.11	6.61	10.24
DER	8.09	7.31	8.99	12.66	5.19	4.59	6.43	10.49
Con-R	7.20	6.50	8.04	9.73	4.59	3.94	4.83	6.39
RankSim	7.02	6.49	7.84	9.68	4.53	4.13	5.37	6.89
VIR	6.99	**6.39**	7.47	**9.51**	4.41	4.07	5.05	**6.23**
LDS+FDS+DER	8.18	7.44	9.52	11.45	5.30	4.75	6.74	7.68
Ours	**6.87**	6.54	**6.96**	9.83	**4.30**	4.10	**4.39**	6.45

Table 1: Evaluation on AgeDB-DIR.

Method \ Shot	MAE↓				GM↓			
	All	Many.	Med.	Few.	All	Many.	Med.	Few.
VANILLA	8.06	7.23	15.12	26.33	4.57	4.17	10.59	20.46
SMOTER	8.14	7.42	14.15	25.28	4.64	4.30	9.05	19.46
SMOGN	8.03	7.30	14.02	25.93	4.63	4.30	8.74	20.12
SMOGN + LDS	8.02	7.39	13.71	23.22	4.63	4.39	8.71	15.80
RRT+LDS	7.79	7.08	13.76	24.64	4.34	4.02	8.72	16.92
SQINV+LDS	7.83	7.31	12.43	22.51	4.42	4.19	7.00	13.94
FOCAL-R+LDS	7.90	7.10	14.72	25.84	4.47	4.09	10.11	19.14
BMC	8.08	7.52	12.47	23.29	-	-	-	-
GAI	8.12	7.58	12.27	23.05	-	-	-	-
VAE	8.04	7.20	15.05	26.30	4.57	4.22	10.56	20.72
DER	7.85	7.18	13.35	24.12	4.47	4.18	8.18	15.18
Con-R	7.33	6.75	11.99	22.22	4.02	3.79	6.98	12.95
RankSim	7.50	6.93	12.09	21.68	4.19	3.97	6.65	13.28
VIR	**7.19**	**6.56**	11.81	20.96	**3.85**	**3.63**	6.51	12.23
LDS + FDS + DER	7.24	6.64	11.87	23.44	3.93	3.69	6.64	16.00
Ours	7.22	6.71	**11.42**	**20.25**	3.88	3.68	**5.74**	**11.13**

Table 2: Evaluation on IMDB-WIKI-DIR.

with the smoothing-based methods (e.g., as VIR (Wang and Wang 2023; Yang et al. 2021)).

Ablation Study on Group Numbers

We also provide a detailed ablation study on the group numbers with the group prediction accuracy and the MAE on AgeDB in Fig.7 and Fig.8. With the increasing of the group numbers, the prediction accuracy gradually drops, the reason why this phenomenon occurs comes from the data dependence over the groups. Therefore, we proposed soft labeling which can leverage the data dependence across the groups and yield a satisfying outcome.

In Fig.8, we can observe that each portion of data (Majority, Median, and Minority) varies slightly with the increasing of group numbers. Specifically, in 15, 20, 25, and 40 group settings, the performance of the majority shots is always close to each other while the median is varied slightly. Meanwhile, most of them always outperform other DIR solutions in Tab.1 and Fig.8, which also shows the prominence of our proposed method.

Method \ Shot	MSE↓				Pearson Correlation↑			
	All	Many.	Med.	Few.	All	Many.	Med.	Few.
VANILLA	0.974	0.851	1.520	0.984	74.2	72.0	62.7	75.2
SMOTER	1.046	0.924	1.542	1.154	72.6	69.3	65.3	70.6
SMOGN	0.990	0.896	1.327	1.175	73.2	70.4	65.5	69.2
SMOGN + LDS	0.962	0.880	1.242	1.155	74.0	71.5	65.2	69.8
RRT	0.964	0.842	1.503	0.978	74.5	72.4	62.3	75.4
RRT + LDS	0.916	0.817	1.344	0.945	75.7	73.5	64.1	76.6
FOCAL-R	0.951	0.843	1.425	0.957	74.6	72.3	61.8	76.4
INV	1.005	0.894	1.482	1.046	72.8	70.3	62.5	73.2
INV + LDS	0.914	0.819	1.31	0.95	75.6	73.4	63.8	76.2
VAE	0.968	0.833	1.511	1.102	75.1	72.4	62.1	74.0
LDS + FDS	0.907	0.802	1.363	0.942	76.0	74.0	65.2	76.6
DER	1.001	0.912	1.368	1.055	73.2	71.1	64.6	74.0
RankSim	0.903	0.908	0.911	0.804	75.8	70.6	69.0	82.7
VIR	0.892	**0.795**	0.899	0.781	**77.6**	**75.2**	69.6	84.5
LDS + FDS + DER	1.007	0.880	1.535	1.086	72.9	71.4	63.5	73.1
Ours	**0.887**	0.897	**0.891**	**0.779**	77.4	74.9	**70.7**	**85.8**

Table 3: Evaluation on STS-B-DIR.

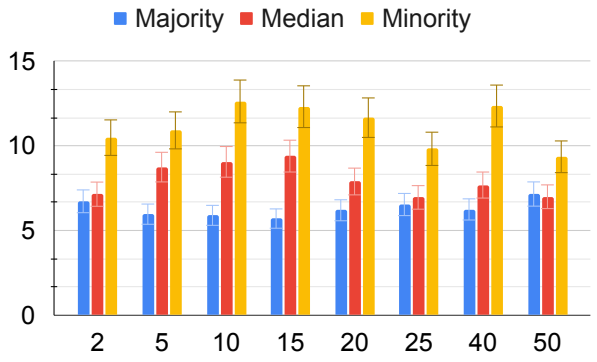

Figure 8: Comparison on Group Numbers **vs** MAE in Majority, Median and Minority. Y: MAE, X: group numbers.

Conclusion

In this work, we present a symmetric descending Soft labeling guided group-aware ordinal contrastive learning framework to learn a high-quality representation that both exhibits discriminative and similar characteristics simultaneously to address the DIR with a multi-expert regressor in a divide-and-conquer manner motivated by our theoretical analysis. Extensive experiments on various real-world datasets verify the superiority of our method. Our analysis of the results further validates the effectiveness of our proposed method.

Acknowledgments

We appreciate constructive feedback from anonymous reviewers and meta-reviewers. Thanks to Dr.Qi Chen for her valuable suggestions. This work is supported by the Natural Sciences and Engineering Research Council of Canada

(NSERC), Discovery Grants program.

References

Branco, P.; Torgo, L.; and Ribeiro, R. P. 2017. SMOGN: a pre-processing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications*, 36–50. PMLR.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9260–9269.

Gong, Y.; Mori, G.; and Tung, F. 2022. RankSim: Ranking Similarity Regularization for Deep Imbalanced Regression. In *International Conference on Machine Learning (ICML)*.

He, H.; and Garcia, E. A. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9): 1263–1284.

Keramati, M.; Meng, L.; and Evans, R. D. 2023. ConR: Contrastive Regularizer for Deep Imbalanced Regression. *arXiv preprint arXiv:2309.06651*.

Liu, H.; HaoChen, J. Z.; Gaidon, A.; and Ma, T. 2021. Self-supervised Learning is More Robust to Dataset Imbalance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; and Zafeiriou, S. 2017. AgeDB: The First Manually Collected, In-the-Wild Age Database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1997–2005.

Pintea, S. L.; Lin, Y.; Dijkstra, J.; and van Gemert, J. C. 2023. A step towards understanding why classification helps regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19972–19981.

Ren, J.; Zhang, M.; Yu, C.; and Liu, Z. 2022. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7926–7935.

Rothe, R.; Timofte, R.; and Van Gool, L. 2018. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *International Journal of Computer Vision*, 126.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels, Belgium: Association for Computational Linguistics.

Wang, Z.; and Wang, H. 2023. Variational Imbalanced Regression: Fair Uncertainty Quantification via Probabilistic Smoothing. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xiao, Y.; Zhang, L.; Liu, B.; Cai, R.; and Hao, Z. 2023. Multi-task ordinal regression with labeled and unlabeled data. *Information Sciences*, 649: 119669.

Yang, Y.; Zha, K.; Chen, Y.; Wang, H.; and Katabi, D. 2021. Delving into Deep Imbalanced Regression. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 11842–11851. PMLR.

Zha, K.; Cao, P.; Son, J.; Yang, Y.; and Katabi, D. 2023a. Rank-N-Contrast: Learning Continuous Representations for Regression. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zha, K.; Cao, P.; Son, J.; Yang, Y.; and Katabi, D. 2023b. Rank-N-Contrast: Learning Continuous Representations for Regression. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhang, S.; Yang, L.; Mi, M. B.; Zheng, X.; and Yao, A. 2023. Improving Deep Regression with Ordinal Entropy. In *The Eleventh International Conference on Learning Representations*.