Imbalanced Label Distribution Learning

Xingyu Zhao1,2, Yuexuan An1,2, Ning Xu1,2, Jing Wang1,2, Xin Geng1,2*

1School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
2Key Laboratory of Computer Network and Information Integration (Ministry of Education), Southeast University, Nanjing 211189, China
{xyzhao, yx_an, xning, wangjing91, xgeng}@seu.edu.cn

Abstract

Label distribution covers a certain number of labels, representing the degree to which each label describes an instance. The learning process on the instances labeled by label distributions is called Label Distribution Learning (LDL). Although LDL has been applied successfully to many practical applications, one problem with existing LDL methods is that they are limited to data with balanced label information. However, annotation information in real-world data often exhibits imbalanced distributions, which significantly degrades the performance of existing methods. In this paper, we investigate the Imbalanced Label Distribution Learning (ILDL) problem. To handle this challenging problem, we delve into the characteristics of ILDL and empirically find that the representation distribution shift is the underlying reason for the performance degradation of existing methods. Inspired by this finding, we present a novel method named Representation Distribution Alignment (RDA). RDA aligns the distributions of feature representations and label representations to alleviate the impact of the distribution gap between the training set and the test set caused by the imbalance issue. Extensive experiments verify the superior performance of RDA. Our work fills the gap in benchmarks and techniques for practical ILDL problems.

Introduction

Learning with ambiguity is one of the most important machine learning topics since data ambiguity is ubiquitous in the real world (Geng 2016; Gao et al. 2017). Label distribution learning (LDL) is a novel paradigm for dealing with data ambiguity. LDL assigns each instance a label distribution and learns the mapping from instances to label distributions. Each element of a label distribution is called the label description degree that explicitly indicates the relative importance of the corresponding label to an instance. As the utility of dealing with ambiguity explicitly, LDL has been successfully applied to many real applications, such as facial landmark detection (Su and Geng 2019), age estimation (Gao et al. 2018), head poses estimation (Geng and Xia 2014), zero-shot learning (Huo and Geng 2017), emotion analysis (Yang et al. 2021a) and autism spectrum disorder classification (Wang et al. 2022).

Despite the fact that LDL achieved success in many applications, one limitation with existing LDL methods is that they are designed for data with balanced supervision information in different labels. That is, the distribution of label annotation information is balanced, which means the sum of the label description degree for each label is approximately equal. However, annotation information in real-world data often exhibits imbalanced distributions, which significantly degrades the performance of existing methods (He and Garcia 2009; Wu et al. 2020). For example, when training a movie rating distribution model for some types of movies, the description degree distribution corresponding to a certain rating may be much higher than other ratings due to the possible deviation of the data collection means. Therefore, the ideal dataset shown in Figure 1(a) may be difficult to gather, and it is possible to obtain the imbalanced dataset shown in Figure 1(b). We use a standard LDL method (i.e., SA-BFGS) to train two models from these training sets separately and test them on the given balanced test set (Geng 2016). From the results given in Figure 1(c), we can find that the performance of the model trained on the imbalanced dataset is significantly worse than that of the model trained on the balanced dataset on each evaluation criterion. Therefore, how to learn an LDL model resilient to the imbalanced label distribution is challenging and meaningful for the practicality of LDL.

We refer to this new and challenging scenario as Imbalanced Label Distribution Learning (ILDL) and systemati-
We identify Imbalance Label Distribution Learning (ILDL) as a new challenging topic and formally define the setting of the ILDL problem.

We delve into the characteristics of the ILDL problem and empirically find that the representation distribution shift is the underlying reason for the performance degradation of existing methods.

We propose a novel method named Representation Distribution Alignment (RDA) for the ILDL problem based on our findings.

We set up three strong baseline methods for the ILDL problem by reshaping the objective functions of existing imbalanced learning approaches.

We curate several benchmark datasets for proper ILDL performance evaluation.

We propose a novel method named Representation Distribution Alignment (RDA) for the ILDL problem based on our findings.

We set up three strong baseline methods for the ILDL problem by reshaping the objective functions of existing imbalanced learning approaches.

We curate several benchmark datasets for proper ILDL performance evaluation.

Related Work

Label Distribution Learning

Label distribution learning (LDL) is a novel learning paradigm, which assigns an instance a label distribution and learns a mapping from instances to label distributions straightly (Geng 2016). In recent years, LDL has been widely studied. (Geng, Yin, and Zhou 2013) proposes the first specialized LDL algorithm, whose objective function consists of the maximum entropy model (Berger, Pietra, and Pietra 1996) and KL divergence. (Zhao and Zhou 2018) casts the label correlations exploration as a ground metric learning problem and adopts optimal transport distance to measure the quality of prediction. (Ren et al. 2019a) exploits the label correlations and learns the common features for all labels and specific features for each label simultaneously. (Jia et al. 2019) exploits local label correlation by capturing low-rank structure on clusters of samples with trace-norm regularization. (Zheng, Jia, and Li 2018) and (Jia et al. 2021) consider label correlation to be local and learn optimal encoding vector and label distribution simultaneously. (Ren et al. 2019b) captures global label correlation with a low-rank matrix and updates the matrix on different clusters to explore local label correlation, which exploits both global and local label correlations. However, these methods are not resilient to the ILDL problem since they do not consider the distribution gap between the training set and the test set.

Imbalanced Learning

Arising from long-tail distributions of natural data, imbalanced learning has been extensively studied. Imbalanced classification (also referred to as long-tailed recognition) (Liu et al. 2019) is one popular topic and numerous methods have been proposed. These works mainly follow two directions. One line of these approaches is re-sampling, which uses under-sampling (Buda, Maki, and Mazurowski 2018)
or over-sampling (Byrd and Lipton 2019) to achieve a relatively balanced dataset. However, the former might weaken feature learning capacity due to omitting a number of valuable samples, and the latter might lead to over-fitting minority classes with duplicated samples. In the meantime, (Wu et al. 2020) indicates that the adaption of this technology into the multi-label setting will not cause a significant change in the label frequency. The other line of these approaches is cost-sensitive, which assigns a weight to each sample according to cost metrics. (Lin et al. 2020) uses the output of the predictive model as the weights. (Cui et al. 2019) proposes a novel class-balanced loss that re-weighting the loss of different labels by the inverse of the effective number of samples. (Wu et al. 2020) applies re-weighting based on the class frequency and modifies the loss gradient with a regularization as well for better optimization.

There are also several works that focus on imbalanced regression. (Torgo et al. 2013) is the first work to address this problem by adopting the SMOTE algorithm (Chawla et al. 2002). (Branco, Torgo, and Ribeiro 2017) presents a Gaussian noise-based synthetic case generation method. (Branco, Torgo, and Ribeiro 2018) introduces a bagging-based ensemble method. Recently, (Yang et al. 2021b) further delves into this problem. It exploits the similarity between nearby targets in target space and feature space, and proposed two smoothing methods for targets and features.

Intuitively, ILDL is similar to the existing imbalanced classification and regression problems in that specific target values have significantly fewer observations (Liu et al. 2019; Cao et al. 2019; Zhou et al. 2020; Yang et al. 2021b). However, it brings greater challenges distinct from imbalanced classification and regression. Compared with imbalanced classification, its target values for each label become continuous, which causes ambiguity when directly applying existing approaches such as re-sampling and re-weighting. Compared with imbalanced regression, ILDL considers not only the continuous target values but also the distribution of these values brought by multiple related label dimensions.

**Problem Setting**

First, the main notations used in this paper are listed as follows. The instance variable is denoted by \( x \), the particular \( i \)-th instance is denoted by \( x_i \), the label variable is denoted by \( y \), the particular \( j \)-th label value is denoted by \( y_{ij} \), the description degree of \( y \) to \( x \) is denoted by \( d_{xy} \), the label distribution of \( x_i \) is denoted by \( d_i = (d_{y_{1|i}}, d_{y_{2|i}}, \ldots, d_{y_{c|i}}) \), where \( c \) is the number of possible label values, \( d_{y_{ci}} \in [0, 1] \) and \( \sum_{j=1}^{c} d_{y_{ji}} = 1 \). In this paper, we consider the imbalanced label distribution setting where the sum of description degrees of different labels are significantly different. Formally, we define the imbalanced label distribution learning problem as follows.

**Problem 1 (Imbalanced Label Distribution Learning, ILDL).** Let \( \mathcal{X} = \mathbb{R}^q \) denote the input space and \( \mathcal{Y} = \{y_1, y_2, \ldots, y_c\} \) denote the complete set of labels. We consider an imbalance training set \( S = \{(x_i, d_i)\}_{i=1}^{N} \), where \( x_i \in \mathcal{X} \) is a \( q \)-dimensional real value vector, \( d_i = (d_{y_{1|i}}, d_{y_{2|i}}, \ldots, d_{y_{c|i}})^T \in [0, 1]^c \) is the corresponding \( c \)-dimensional label distribution. The imbalanced factor of the training set \( \gamma = \max \left\{ \left( \sum_{i=1}^{N} d_{y_{ij}} \right)_{j=1}^{c} \right\} / \min \left\{ \left( \sum_{i=1}^{N} d_{y_{ij}} \right)_{j=1}^{c} \right\} > 10 \). The goal of ILDL is to learn a mapping from an instance \( x \) to its corresponding label distribution \( \mathbf{d} \), which can achieve high performance on a balanced test set \( S^* = \{(x_i^*, d_i^*)\}_{i=1}^{N} \).

Our purpose is to train an LDL model on the imbalanced training set, which can achieve better performance in the relatively balanced test set.

**Methods**

In this section, we first reshape several objective functions of existing imbalanced learning approaches as strong baselines for the ILDL problem. Furthermore, we propose a novel method, **Representation Distribution Alignment**, which alleviates the impact of the distribution gap between the training set and the test set by aligning the distributions of feature representations and label representations of instances.

**Objective Function Reshaping**

In LDL, Kullback-Leibler (KL) divergence between the ground truth and the predicted label distribution is a commonly used loss function. Assume \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]

where \( f_\theta(\cdot) \) is a mapping from \( \mathcal{X} \) to \( \mathcal{Y} \). The objective function of LDL can be formulated

\[
\text{min} \sum_{i=1}^{N} f_\theta(x_i) d_i,
\]
by
\[
\min_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{c} d_{x,i}^{v_j} \ln \frac{d_{x,i}^{v_j}}{f_{\theta}^{(j)}(x_i)},
\]
(1)
where \( f_{\theta}^{(j)}(x_i) \) is the output for the \( j \)-th label of \( f_{\theta}(x_i) \). The plain KL divergence loss function may be vulnerable to label imbalance due to the observations of different labels are significantly different. Therefore, we reshape the objective functions from imbalanced classification to make it be resilient to ILDL.

**Focal loss.** Focal loss places a higher weight of loss on instances predicted with low probability on ground truth to emphasize the importance of “hard-to-classify” instances. (Lin et al. 2020). In ILDL, we modify the original focal loss to the following form:

\[
\mathcal{L}_{OFR-FL} = \sum_{i=1}^{n} \sum_{j=1}^{c} (1 - f_{\theta}^{(j)}(x_i)) \gamma d_{x,i}^{v_j} \ln \frac{d_{x,i}^{v_j}}{f_{\theta}^{(j)}(x_i)},
\]
(2)
where \( \gamma \geq 0 \) is a tunable focusing parameter. The idea of Eq. (2) is consistent with the original focal loss, which utilizes the predicted values of different labels to weight the original loss, so as to better deal with label imbalance.

**Class-balanced focal loss.** Class-balanced focal loss estimates the effective number of samples of each class and uses them to further reweight focal loss (Cui et al. 2019). Compared with focal loss, class-balanced focal loss integrates the class-level information into the loss function, which can capture the diminishing marginal benefits of data and reduce redundant information of head classes. In ILDL, we modify the number of samples of each class to the sum of the description degree of each label, i.e., \( \hat{N}_j = \sum_{i=1}^{N} d_{x,i}^{v_j} \), and the “effective number” of samples of each class is re-defined as the “effective description degree” of each label, i.e., \( r_{CB}^{(j)} = 1 - \beta/(1 - \beta^{N_j}) \). The modified class-balanced focal loss is defined as

\[
\mathcal{L}_{OFR-CB} = \sum_{i=1}^{n} \sum_{j=1}^{c} r_{CB}^{(j)} (1 - f_{\theta}^{(j)}(x_i)) \gamma d_{x,i}^{v_j} \ln \frac{d_{x,i}^{v_j}}{f_{\theta}^{(j)}(x_i)}.
\]
(3)

**Distribution-balanced focal loss.** Distribution-balanced loss (Wu et al. 2020) is first proposed for multi-label classification. It consists of re-balanced weighting and negative-tolerant regularization. Re-balanced weighting assigns different weights to each label for each sample based on the re-sampling strategy. In ILDL, we modify the re-balancing weight to \( r_{i}^{(j)} = P_{C}^{(j)}(x_i)/P_{I}^{(j)}(x_i) \), where \( P_{C}^{(j)}(x_i) = \frac{1}{N_j} \) is the expectation of label-level sampling frequency and \( P_{I}^{(j)}(x_i) = \frac{1}{2} \sum_{j=1}^{c} d_{x,i}^{v_j} / \hat{N}_j \) is the expectation of instance-level sampling frequency. We also use the smoothing function (Wu et al. 2020) to map \( r \) to a smoothed value \( \tilde{r} \). Negative-tolerant regularization (NTR) tries to deal with the issue that the gradients of the positive classes and the negative classes are significantly different. In ILDL, we modify the predicted value after negative-tolerant regularization to \( q_{x,i}^{v_j} = \exp \left( f_{\theta}^{(j)}(x_i) - v_j \right) / \sum_{k=1}^{c} \exp \left( f_{\theta}^{(k)}(x_i) - v_k \right) \) where \( v \) is a class-specific bias. Combine re-balanced weighting with negative-tolerant regularization, we have the modified distribution-balanced focal loss:

\[
\mathcal{L}_{OFR-DB} = \sum_{i=1}^{n} \sum_{j=1}^{c} \frac{1}{c} \left[ r_{i}^{(j)} (1 - f_{\theta}^{(j)}(x_i)) \gamma \right] \left( 1 - \frac{1}{\lambda} \right) d_{x,i}^{v_j} + \frac{1}{\lambda} \ln \left( \frac{d_{x,i}^{v_j}}{q_{x,i}^{v_j}} \right).
\]

where \( \lambda \) is a balanced hyperparameter.

**Representation Distribution Alignment for ILDL.**

In the former subsection, we reshape several objective functions commonly used in imbalanced classification. These approaches, however, are all independently inside each label, and there is no insight into the distribution of the whole label
set and no interaction among different labels. At the same time, these approaches only pay attention to the processing of the label space, and do not effectively utilize the information of the feature space to improve the performance of the predictive model. As a result, these methods still cannot fully avoid the impact of the representation distribution shift problem.

To tackle these issues, we propose Representation Distribution Alignment (RDA) for ILDL. RDA aligns the distributions of feature representations and label representations of instances to effectively leverage the information hidden in both feature space and label space and alleviate the impact of the distribution gap between the training set and the test set. Specifically, as shown in Figure 4, it first utilizes two mapping functions \( g_r \) and \( h_\phi \) to map the feature \( x \) and label distribution vector \( d \) into latent spaces. Then, it maps the latent vectors \( g_r(x) \) and \( h_\phi(d) \) into a common space and aligns the distributions of feature representations and label representations of instances in this space. Assuming the distributions of features and labels are Gaussian, i.e., there are two mapping functions \( G_r \) and \( H_\phi \) which maps \( g_r(x) \) to \( N(\mu_F, \sigma^2_F) \) and maps \( h_\phi(d) \) to \( N(\mu_L, \sigma^2_L) \), respectively. Then RDA minimizes the differences between \( N(\mu_F, \sigma^2_F) \) and \( N(\mu_L, \sigma^2_L) \). In order to better utilize the learned knowledge for description degree prediction, RDA also adopts a decoding function \( F_\theta \) to decode the feature encoding value to the description degree. Formally, RDA optimizes the following objective function:

\[
\min_{\epsilon, \varphi, \theta} V(\epsilon, \theta) + \lambda_1 \Omega_1(\varphi, \epsilon, \theta) + \lambda_2 \Omega_2(\epsilon, \varphi) + \lambda_3 \Omega_3(\epsilon, \epsilon, \varphi),
\]

where \( V(\epsilon, \theta) \) is the loss function for description degree prediction, \( \Omega_1 \) and \( \Omega_2 \) are used to learn reasonable label representation mapping and feature representation mapping, respectively, \( \Omega_3 \) is used for aligning the distributions of feature representations and label representations, \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) are balanced parameters.

The purpose of \( V \) is to reduce the divergence between the real distribution \( d \) and predicted distribution \( F_\theta(g_r(x)) \). Any reshaped objective function can be leveraged as the loss function \( V \). The goal of \( \Omega_1 \) is to make \( h_\phi, G_r \) and \( F_\theta \) have better label representation ability. To achieve that, we first sample values for label representations \( r_l \) using the reparameterization trick (Rezende, Mohamed, and Wierstra 2014); \( r_L = \mu_L + \sigma_L \delta_L \), where \( \mu_L \) and \( \sigma_L \) are computed from \( H_\phi(h_\phi(d)) \) and \( \delta_L \sim N(0, I) \). Then we define \( \Omega_1 \) as the following form:

\[
\Omega_1(\varphi, \epsilon, \theta) = \sum_{i=1}^{n} \sum_{j=1}^{c} KLD_{e^i_j} = \sum_{i=1}^{c} \ln \frac{d_{e^i}}{F_\theta(r_L)}.
\]

For \( \Omega_2 \), we leverage the information of label space to promote the feature representation ability of the model. Specifically, we optimize the KL divergence between \( N(\mu_F, \sigma^2_F) \) and \( N(\mu_L, \sigma^2_L) \):

\[
\Omega_2(\epsilon, \epsilon) = -\frac{1}{2} \sum_{k=1}^{K} \left[ \log \mu^{(k)} - \mu^{(k)} - \tau^{(k)} + 1 \right].
\]

\( K \) is the dimension of the latent space, \( \tau^{(k)} \) denotes the k-th element, \( \mu^{(k)} = \frac{\mu_f^{(k)}}{\mu_L^{(k)}} \), \( \tau^{(k)} = \frac{(\mu_f^{(k)} - \mu_L^{(k)})^2}{\mu_L^{(k)}} \), \( \mu_f \) and \( \mu_L \) are computed from \( g_r(x) \) and \( h_\phi(d) \).

The aim of \( \Omega_3 \) is to align the distributions of feature representations and label representations. Specifically, we align the similarities of the distributions of both representations. For the feature representations \( r_f \), we again use the reparameterization trick : \( r_f = \mu_f + \sigma_f \delta_f \), where \( \delta_f \sim N(0, I) \). Then the similarity matrix \( A \) of the distribution of feature representations can be obtained by:

\[
A_{mn} = S(r_f^m, r_f^n),
\]

where \( S \) is cosine similarity, \( r_f^m \) and \( r_f^n \) are the m-th and n-th instances. Meanwhile, the similarity matrix \( Z \) of the distribution of label representations can be obtained by:

\[
Z_{mn} = S(r_L^m, r_L^n). 
\]

For \( \Omega_3 \), the distance between \( A \) and \( Z \) is minimized:

\[
\Omega_3(\epsilon, \epsilon, \mu, \varphi) = \frac{1}{M^2} \sum_{m=1}^{M} \sum_{n=1}^{M} (A_{mn} - Z_{mn})^2 
\]

In the training stage, the gradient-based method is used to optimize \( \Omega_3 \). In the prediction stage, given an instance \( x^* \), the prediction of RDA can be obtained by \( F_\theta(g_r(x^*)) \).

**Experiments**

**Datasets**

We curate six ILDL benchmarks that span movie rating, facial beauty perception and visual sentiment distribution perception. These datasets are sampled from six standard LDL datasets, including Movie (Geng 2016), SCUT-FBP (Xie et al. 2015), Emotion6 (Peng et al. 2015), FlickrLDL (Yang, Sun, and Sun 2017), RAF-ML (Shang and Deng 2019) and Natural Scene (Geng 2016). The sampling process is performed 10 times, for each time, we sample the training set.
and validation set from the original training set which occupies 90% of the examples, while the test set remains unchanged. Figure 5 illustrates the distribution of label annotation information of these datasets and their level of imbalance.

### Evaluation Criteria

Six standard LDL measures (Chebyshev Distance, Clark Distance, Canberra Metric, Kullback-Leibler Divergence, Cosine Coefficient, and Intersection Similarity between ground-truth label distributions and predicted label distributions) are selected to evaluate different methods for the prediction of label distributions. Besides, Euclidean Distance is also adopted to evaluate the performance of different methods on tail, head and all labels. It is worth noting that the evaluation criteria of ILDL are different from imbalanced classification and regression problems. In ILDL, the description degrees of the tail labels in the training set tend to increase in the test set, while the description degrees of the head labels tend to decrease. Therefore, head labels and tail labels are equally important in ILDL.

### Methodology

Several existing state-of-the-art LDL algorithms, i.e., SA-BFGS (Geng 2016), EDL-LRL (Jia et al. 2019), LDLSF (Ren et al. 2019a), LDL-LCLR (Ren et al. 2019b), Adam-LDL-SCL (Jia et al. 2021) and LDL-LDM (Wang and Geng 2021) are set as baselines. Three objective function reshaping approaches, i.e., OFR-FL, OFR-CB and OFR-DB, are also performed in the experiments. Moreover, our RDA is compared with these methods. In RDA, \( g_i \), \( h_\phi \) and \( F_\theta \) are set as linear projections, \( G_c \) and \( H_c \) are set as single-layer neural network with two outputs including mean and variance of Gaussian, and the modified distribution-balanced focal loss is adopted as the loss function \( V \). Hyperparameters \( \lambda_1 \), \( \lambda_2 \) and \( \lambda_3 \) are selected by grid search from the set \{0.01, 0.05, 0.1, 0.2, 0.5\}.

### Main Results

#### Comparisons on Distribution Criteria

Tables 1 and 2 tabulate the experimental results of different methods on Chebyshev Distance and Kullback-Leibler Divergence. For each evaluation criterion, ‘−’ indicates the smaller the better. In Tables 1 and 2, the two-tailed t-test at 0.05 significance level is conducted, and the best performances are highlighted in bold. ‘/’ indicates whether RDA is statistically superior/inferior to the comparing methods. From Tables 1 and 2, it can be observed that: 1) The existing LDL methods show poor performances in solving ILDL tasks. 2) The performances of the objective function reshaping approaches are superior to existing methods. 3) Compared with these baseline methods, our RDA has achieved better results on Chebyshev Distance and Kullback-Leibler Divergence and significantly outperforms other algorithms in most cases. These observations indicate that the RDA can effectively alleviate the impact of the distribution gap between the training set and the test set.

#### Comparisons on Imbalance Criteria

Table 3 gives the experimental evaluations of different algorithms on tail, head and all labels. From Table 3, we can find that the most objective function reshaping approaches have better performance than existing LDL methods. In the meantime, RDA achieves the best performance in almost all cases. In particular, RDA achieved the best performance on head and all labels in all...
cases. These observations demonstrate the effectiveness of the proposed methods in tackling ILDL tasks.

**Effect of Hyperparameters $\lambda_1$, $\lambda_2$ and $\lambda_3$**

In this subsection, we explore the effect of hyperparameters $\lambda_1$, $\lambda_2$ and $\lambda_3$. We compare the performances of RDA with different values of $\lambda_1$, $\lambda_2$ and $\lambda_3$ on the six datasets measured by Chebyshev Distance. Figure 6 illustrates the performances of RDA with different values of $\lambda_1$, $\lambda_2$ and $\lambda_3$. From these curves, we can find that: 1) Overall, RDA has stable performances with a wide range of hyperparameter values on all six datasets; 2) Appropriate values of $\lambda_2$ can bring slight performance gains on some datasets; 3) The performance of the model hardly changes with changes in $\lambda_1$ and $\lambda_3$. These findings further demonstrate the robustness of the proposed RDA.

**Further Analyses**

We compare the average ranks of different algorithms over all the six ILDL datasets and find that RDA surpasses the compared methods by a significant margin across all the evaluation criteria, which further indicates the effectiveness of the proposed RDA. Details of the further analyses are provided in the appendix, which is available at: https://github.com/ailearn-ml/RDA.

**Conclusion**

We study a challenging and meaningful problem, i.e., Imbalanced Label Distribution Learning (ILDL), in this paper. We curate several benchmark ILDL datasets and offer three strong baselines. Moreover, we delve into the characteristics of the ILDL problem and find that the representation distribution shift is the underlying reason for the performance degradation of existing methods. Based on this finding, we propose a novel method named Representation Distribution Alignment, which can align the distributions of feature representations and label representations to effectively alleviate the impact of the distribution gap between the training set and the test set caused by the imbalance issue. Extensive experiments confirm the superior performance of our proposed method. Our work fills the gap in benchmarks and techniques for practical ILDL problems.
Acknowledgements

This research was supported by the National Key Research & Development Plan of China (No. 2018AAA0100104), the National Science Foundation of China (62125602, 62076063, 62206050), China Postdoctoral Science Foundation (2021M700023, 2022M720028), Jiangsu Province Science Foundation for Youths (BK20210220), Young Elite Scientists Sponsorship Program of Jiangsu Association for Science and Technology (TJ-2022-078).

References


