ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448034

PSEUDO LABELS REGULARIZATION FOR IMBALANCED PARTIAL-LABEL LEARNING

Mingyu Xu[1,2] *Zheng Lian*[1,2] *Bin Liu*[1,2*] *Zerui Chen*[3] *Jianhua Tao*[4]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, 100190, China
[2] Institute of Automation, Chinese Academy of Sciences, 100190, China
[3] Chinese-American Joint Program of RDFZ XISHAN School, 100193, China
[4] Department of Automation,Tsinghua University, 100084, China

ABSTRACT

Partial-label learning (PLL) is an important branch of weakly supervised learning where the single ground truth resides in a set of candidate labels, while the research rarely considers the label imbalance. A recent study for imbalanced PLL propose that the combinatorial challenge of partial-label learning and long-tail learning lies in matching between a decent marginal prior distribution with drawing the pseudo labels. However, even if the pseudo label matches the prior distribution, the tail classes will still be difficult to learn because the total weight of tail classes is too small. Therefore, we propose a pseudo-label regularization technique specially designed for imbalanced PLL. By punishing the pseudo labels of head classes, our method implements state-of-art under the standardized benchmarks compared to the previous PLL methods.

Index Terms— Partial label learning, Pseudo label, Imbalanced learning

1. INTRODUCTION

In the real world, a large number of data are crowdsourced to non-experts for annotation. Due to the existence of various noises, the annotators may hesitate in some labels [1, 2, 3] and give a candidate label set for an instance. To deal with the problem, PLL has attracted significant attention from the community [4, 5, 6]. PLL is an important branch of weakly supervised learning, which assumes the single ground-truth label must be in the candidate set. A plethora of methods have been developed to tackle this problem, including average-based method [5, 7], identification-based [8, 9, 10, 11, 12, 13].

However, existing PLL methods usually considered on the data set of balanced category, which may not hold in practice. In many real world scenarios, training data exhibits a long-tailed label distribution. That is, many labels occur infrequently in the training data [14]. Unbalanced data will cause predicted values to deviate from tail categories [15, 16]. The consideration of imbalanced categories is meaningful for the wider application of partial-label learning. Recently, researchers propose a novel framework for long-tailed partial-label learning, called Solar [17], which proposes that the combinatorial challenge of partial-labeling and long-tail learning lies in matching between a decent marginal prior distribution with drawing the pseudo labels. However, we find that there is still a good result in the case of mismatching in their experiment, which makes us wonder whether matching is necessary. On the other hand, the calculation of pseudo label in Solar requires Sinkhorn-Knopp algorithm [18]. It is an iterative algorithm that consumes more time and space. More importantly, its convergence does not exist under the setting of PLL. For this reason, Solar also needs to use the relaxed solution occasionally, which inspired us to propose a simpler yet more effective algorithm.

Therefore, we propose a pseudo-label regularization technique for imbalanced partial-label learning. We do not require the matching between the pseudo labels and distribution matching and the closed-form solution of the pseudo labels can be obtained directly by using the Lagrange multiplier method. We comprehensively evaluate our method on various benchmark datasets, where our method establishes state-of-the-art performance. Compared to Solar, our method improves the average accuracy by 4.53% on the long-tailed version of CIFAR-10 and 2.35% on the long-tailed version of CIFAR-100.

2. METHOD

2.1. Problem Setup

Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, 2, \cdots, c\}$ be the label space with c distinct categories. We consider a partially labeled dataset $\mathcal{D} = \{(x_i, S(x_i))\}_{i=1}^{N}$ where N is the number of samples and $S(x_i) \in \{0, 1\}^c$ is the candidate set for the sample $x_i \in \mathcal{X}$. We denote the j^{th} element of $S(x_i)$ as $S_j(x_i)$. Here, $S_j(x_i)$ is equal to 1 if the label j is a candidate label for x_i, and otherwise 0.

Our goal is to train a classifier $f : \mathcal{X} \mapsto [0, 1]^c$, parameterized by θ. Here, f is the softmax output of a neural network, and $f_j(\cdot)$ denotes the j^{th} entry. To perform label disambiguation, we maintain a pseudo-label $w(x_i)$ for sample x_i, where $w_j(x_i)$ donate the j^{th} entry. We train the classifier with the cross-entropy loss $\sum_{j=1}^{x} -w_j(x_i) log f_j(x_i)$. For the convenience of writing, also record $w_j(x_i)$ as w_{ij}, $f_j(x_i)$ as f_{ij}, $S_j(x_i)$ as S_{ij}. And we noted the proportion of each category as $r \in \mathcal{R}^c$.

2.2. Motivation

PRODEN [9] is the classic way to use pseudo labels for PLL, which does not take into account the imbalance but is able to write closed-form solution as:

$$w_{ij} = \frac{S_{ij} f_{ij}}{\sum_{j=1}^{c} S_{ij} f_{ij}} \tag{1}$$

Recently, researchers have proposed a method that can conduct long tail PLL learning, named Solar[17]. Solar adopts the Sinkhorn-Knopp algorithm to obtain pseudo label w iteratively, which is a special case of the dual gradient ascent method. However, it is difficult to guarantee the convergence of Sinkhorn-Knopp algorithm when we

*Corresponding Author. This work is supported by the National Natural Science Foundation of China (NSFC) (No.62276259, No.62201572, No.U21B2010, No.62271083, No.62306316)

Algorithm 1 Pseudo-code of Pseudo Labels Regularization for Imbalanced Partial-Label Learning

Input: Training dataset $\mathcal{D}$, classifier f, uniform marginal r,hyperparameters λ, M.

for epoch = 1,2,... **do**
  for step = 1,2,... **do**
    Get classifier prediction f_{ij} on a mini-batch of data B.
    Get the pseudo labels by Equaion 5.
    Select sample with small loss in each category.
    Calculate classification loss, consistency loss and mixup loss.
    Update SGD optimizer to update f.
  end for
  Update the marginal r with the information of the training set.
end for

solve the problem. So the relaxed solution is used in Solar. In order to avoid the discussion of convergence and save time and space, can we propose a method that can directly write the closed-form solution like PRODEN for imbalanced PLL? Besides, we note that in Solar, the imprecise piror r and real r can lead to similar performance, which indicates that even if there is no match between imprecise r and real r, we may also achieve good results in imbalance PLL as long as we punish the pseudo labels of head classes. Based on the above ideas, we propose a pseudo label regularization method designed for imbalanced biased label learning.

2.3. Pseudo labels regularization

In this section, we describe our novel framework for partial-label learning. This framework will meet the two core points we mentioned above, one is the method that can directly write the closed-form solution, the other is that the method can punish the pseudo labels of head classes. We formalize our method as:

$$\min_{w,\theta} \sum_{i=1}^{N}\sum_{j=1}^{c}(-w_{ij}logf_{ij} + \frac{1}{\lambda}w_{ij}logw_{ij} + \frac{M}{\lambda}w_{ij}logr_j) \quad (2)$$

$$s.t. w_{ij} = 0 \; for \; S_{ij} = 0, i \in \{1, ..., N\}, j \in \{1, ..., c\} \quad (3)$$

$$\sum_{j=1}^{c} w_{ij} = 1, i \in \{1, ..., N\} \quad (4)$$

where $\lambda, M > 0$. The objective function consists of three items. The first item is the commonly used cross-entropy classification loss. The second item is about the entropy regularization item of the pseudo labels w to avoid overconfidence. The third item is about the regularization item of the category, which will keep the pseudo labels w away from the prior distribution r.

Similar to PRODEN, we adopt the method of alternately optimizing w and θ. For constraint $w_{ij} = 0 \; for \; S_{ij} = 0$, we can delete the items related to w_{ij} in the optimization goal if $S_{ij} = 0$. Then we can calculate the Hession matrix of the optimization objective as $\frac{1}{\lambda}diag(\{w_{ij}\}_{i\in\{1,...,N\},j\in\{1,...,c\},S_{ij}=1})$, which is a positive definite matrix. This means that the objective function is a convex optimization problem with respect to w and there is a unique optimal solution about w in the case of fixed θ.

Using Lagrange multiplier method, we will know the optimal w satisfy:

$$w_{ij} = \frac{S_{ij}f_{ij}^{\lambda}r_j^{-M}}{\sum_{j=1}^{c} S_{ij}f_{ij}^{\lambda}r_j^{-M}} \quad (5)$$

Besides, we use the SGD optimization to update θ.

Because we can directly write the optimal closed-form solution of w in each iteration process, the cost of space and time is very small. And our method is consistent with Proden under the condition of balanced PLL. This is because when $r_1 = r_2 = ... = r_c$, equation 5 degenerate to PRODEN. And when $M > 0$, the pseudo label of head class is punish hardly by big r_j^{-M}. Because head class has many samples, being slightly punished will not affect its performance too much. However, tail class can benefit from equation 5 whose performance will be greatly improved, so as to improve the overall performance. M is a balance factor to balance head classes and tail classes training in training. Another case where our method degenerates to PRODEN is $M = 0$. In addition, when $\lambda = 1$ and $M = 1$, our method can be degenerates to [19].

2.4. Other technologies

Pseudo labels regularization is the core component of our method. However, in order to better conduct the long-tail PLL learning, we also adopted the following technologies.

Estimate the prior distribution. One of the core difference between long-tail learning and long-tail PLL in that the number of each category is unknown, which requires us to estimate r based on the information of training data. To estimated r, we initialize r to be uniformly distributed $[1/c, ..., 1/c]$, and update it by using a moving-average strategy to ensure the stability of updating:

$$r_j \leftarrow \mu r_j + (1-\mu)\frac{1}{n}\sum_{i=1}^{n} 1_{j=argmax_{1\le j\le c}P_{ij}} \quad (6)$$

where $\mu \in [0, 1]$ is a preset scalar. One advantage of this estimation method is that assuming our classifier can fully predict accurately, the estimated r can approach the true r.

Consistency regulation. Using strong and weak data augmentation is an effective method for PLL learning, which has been verified in PiCO [12], CRDPLL[20], and IRNet [21]. Therefore, we also use consistency regulation in long-tail PLL. Specifically, we calculate the cross entropy loss by using the pseudo label of weakly augmented samples and the prediction of strongly augmented samples.

Mixup. Recently, mixup [22] technology has been used in PLL to enhance the robust of PLL in PiCO+ [13] and DALI [23]. It also shows excellent performance in the long-tail PLL in Solar [17]. In order to improve performance, we also adopted the mixup technology. That is to construct a new sample whose input is a linear combination of two samples, and its pseudo label is also a linear combination of the two samples.

Sample selection After we get the pseudo label from equation 2, only some pseudo are actually trustworthy. Therefore, we will select small loss samples from each type of samples to perform consistency loss and mixup, which will improve the representation ability of our method. We first calculate the samples belonging to k^{th} class from the batch B: $B_k = \{(x_i, w_i) \in B | k = argmax_{1\le j\le c} w_{ij}\}$. Then select the small cross entropy loss sample with numbers of $min(|B_k|, [\rho r_k|B|])$ in each B_k, where $\rho \in [0, 1]$ is a threshold hyper-parameter. Sample select can be seen as a way of curriculum learning [24].

3. EXPERIMENTS

3.1. Setup

Datasets. We evaluate our method on Four long-tailed datasets CIFAR10-LT, CIFAR100-LT, CIFAR100-H-LT [25, 26] and CUB200-

Table 1: Accuracy comparisons on CIFAR10-LT and CIFAR100-LT under various flipping probability ψ and imbalance ratio γ. Bold indicates superior results.

methods	CIFAR10-LT					
	$\psi = 0.3$			$\psi = 0.5$		
	$\gamma = 50$	$\gamma = 100$	$\gamma = 200$	$\gamma = 50$	$\gamma = 100$	$\gamma = 200$
MSE	61.13±1.08	52.59±0.48	48.09±0.45	49.61±1.42	43.90±0.77	39.52±0.70
EXP	52.93±3.44	43.59±0.16	42.56±0.44	50.62±3.00	43.69±2.72	41.07±0.62
LWS	44.51±0.03	43.60±0.12	42.33±0.58	24.62±9.67	27.33±1.84	28.74±1.86
VALEN	58.34±1.05	50.20±6.55	46.98±1.24	40.04±1.88	37.10±0.88	36.61±0.57
CC	78.76±0.27	71.86±0.78	63.38±0.79	73.09±0.40	64.88±1.03	54.41±0.85
PRODEN	81.95±0.19	71.09±0.54	63.00±0.54	66.00±3.60	62.17±3.36	54,65±1.00
PiCO	75.42±0.49	67.73±0.64	61.12±0.67	72.33±0.08	63.25±0.64	53.92±1.64
Solar	83.80±0.52	76.64±1.66	67.47±1.05	81.38±2.84	74.16±3.03	62.12±1.64
Ours	**87.25**±0.51	**81.74**±0.53	**74.07**±1.45	**85.86**±1.01	**78.38**±0.37	**65.76**±2.86
methods	CIFAR100-LT					
	$\psi = 0.05$			$\psi = 0.1$		
	$\gamma = 10$	$\gamma = 20$	$\gamma = 50$	$\gamma = 10$	$\gamma = 20$	$\gamma = 50$
MSE	49.92±0.64	43.94±0.86	37.77±0.40	42.99±0.47	37.19±0.72	31.49±0.35
EXP	25.86±0.94	24.84±0.40	23.58±0.47	24.82±1.41	21.27±1.24	19.88±0.43
LWS	48.85±2.16	35.88±1.29	19.22±8.56	6.10±2.05	7.16±2.03	5.15±0.36
VALEN	49.12±0.58	42.05±1.52	35.62±0.43	33.39±0.65	30.67±0.11	24.93±0.87
CC	60.36±0.52	54.33±0.21	45.83±0.31	57.91±0.41	51.09±0.48	41.74±0.41
PRODEN	60.31±0.50	50.39±0.96	42.29±0.44	47.32±0.60	41.82±0.55	35.11±0.08
PiCO	54.05±0.37	46.93±0.65	38.74±0.11	46.49±0.46	39.80±0.34	34.97±0.09
Solar	64.75±0.71	56.47±0.76	46.18±0.85	61.82±0.71	53.03±0.56	40.96±1.01
Ours	**65.83**±0.43	**58.62**±0.61	**48.73**±0.25	**63.89**±0.63	**54.49**±0.64	**45.74**±0.70

Table 2: Ablation results on CIFAR10-LT ($\psi = 0.5, \gamma = 100$) and CIFAR100- LT ($\psi = 0.1, \gamma = 20$).

methods	CIFAR10-LT	CIFAR100-LT
Ours	78.38	54.49
Ours w/o S	64.28	51.06
Ours w/o MU	65.27	51.97
Ours w/o CR	73.45	51.61
Ours w/o MU+CR	45.97	39.83
Solar	74.16	53.03
Solar w/o S	29.61	35.80
Solar w/o MU+CR	44.83	30.88

LT[27]. The training images are randomly removed class-wise to follow a pre-defined imbalance ratio $\gamma = n_1/n_L$, where n_j is the image number of the j^{th} class. For convenience, class indices are sorted based on the class-wise sample size, in descending order with $n_1 \geq n_2 \geq ... \geq n_L$ and $n_1/n_2 = n_2/n_3 = ... = n_{L-1}/n_L$. We then generate partially labeled datasets by manually flipping negative labels $\hat{y} \neq y$ to false-positive labels with probability $\psi = P(\hat{y} \in \mathcal{Y}|\hat{y} \neq y)$, which follows the settings in previous works [17]. The final candidate label set is composed of the ground-truth label and the flipped false-positive labels.

Baselines. We compare our method with seven state-of-the-art PLL methods: 1) PiCO [12] leverages contrastive learning to disambiguate the candidate labels by updating the pseudo-labels with contrastive prototype labels. 2) PRODEN [9] is also a pseudo-labeling method that iteratively updates the latent label distribution by re-normalized classifier prediction. 3) VALEN [28] recovers the latent label distributions by a Bayesian parametrization model. 4) LWS [11] also works in a pseudo-labeling style, which weights the risk function by considering the trade-off between losses on candidate labels and non-candidate ones. 5) CC [29] is a classifier-consistent method that assumes the candidate label set is uniformly sampled. 6) MSE and EXP [10] utilize mean square error and exponential loss as the risk estimators. All the hyper-parameters are searched according to the original papers. 7) Solar [17] is a long-tail PLL method to match pseudo labels with prior.

Implementation details. We use an 18-layer ResNet as the feature backbone. The model is trained for 1000 epochs using a standard SGD optimizer with a momentum of 0.9. The initial learning rate is set as 0.01, and decays by the cosine learning rate schedule. The batch size is 256. These configurations are applied for our method and all baselines for fair comparisons. We devise a pre-estimation training stage, where we run a model for 100/20 epochs (on CIFAR10/100-LT) respectively to obtain a coarse-grained class prior, which is the same with the previous works [17]. After that, we re-initialize the model weights and run with this class prior. We use $\lambda = 3, M = 2$ for CIFAR10-LT, and $\lambda = 1, M = 0.5$ for CIFAR100-LT. The moving-average parameter μ for class prior estimation is set as 0.1/0.05 in the first stage and fixed as 0.01 later. For class-wise reliable sample selection, we linearly ramp up ρ from 0.2 to 0.5/0.6 in the first 50 epochs. For fair comparisons, we equip all the baselines except PiCO with consistence loss and mixup. The mix coefficient of mixup is sampled from $beta(4, 4)$. For all experiments, we report the mean and standard deviation based on 3 independent runs with different random seeds.

3.2. Main result

Our method achieves SOTA results. As shown in Table 1, our method significantly outperforms all the rivals by a notable margin under various settings of imbalance ratio γ and label ambiguity degree ψ. Specifically, on CIFAR10-LT dataset with $\psi = 0.3$ and the imbalance ratio $\gamma = 200$ we improve upon the best baseline

Table 3: Influence of regularization coefficient λ and M on CIFAR10-LT ($\psi = 0.5, \gamma = 100$).

M ($\lambda = 4$)	All	Many	Medium	Few
1	72.04	89.67	59.92	70.59
2	76.15	87.32	77.77	62.82
3	76.61	84.18	79.39	65.32
4	70.62	82.11	69.68	60.4
M ($\lambda = 3$)	All	Many	Medium	Few
1	73.42	88.65	67.96	65.48
1.5	77.49	87.07	78.33	66.89
2	78.38	85.11	78.75	71.16
3	68.55	79.31	71.23	54.29
M($\lambda = 2$)	All	Many	Medium	Few
0.5	71.33	90.00	67.25	58.09
1	77.49	86.66	79.17	66.07
1.5	73.24	82.81	78.87	56.18
2	68.10	75.38	74.00	52.95
M($\lambda = 1$)	All	Many	Medium	Few
0.25	72.94	86.74	71.11	61.59
0.5	67.24	64.99	74.38	59.93
0.75	65.44	67.28	71.00	51.91
1	61.35	56.47	68.38	54.76

Table 4: Performance comparisons of our method and Solar on the fine-grained CUB200-LT dataset and on the CIFAR100-LT dataset with hierarchical labels (CIFAR100-H-LT).

Methods	Dataset	ALL	Many	Medium	Few
Ours	CUB-200	39.98	61.19	44.31	15.79
Solar		38.96	58.93	42.44	17.18
Ours	CIFAR-100-H	58.88	76.06	60.33	40.17
Solar		58.09	76.78	58.26	39.13

by 6.60%. Specifically, on CIFAR100-LT dataset with ψ = 0.5 and the imbalance ratio $\gamma = 50$, we improve upon the best baseline by 4.00%. Our method is superior to previous methods in all cases. Especially with the increase of the imbalance rate, our method can still show excellent performance.

Ablation Studies We do ablation experiments to illustrate the impact of various technologies in Section 3.4, and can further explain the performance of our proposed methods. 1) Sample select or not. Firstly, we experiment with sample select or not. As a comparison, we also conducted the desired ablation experiment on Solar. w/o S means regards all examples as clean samples and does not perform selection. The results are shown in Table 2. We can find that sample select has improved our method. When we do not use sample select, our method improves by 34.67% in CIFAR10-LT and 15.26 % in CIFAR100-LT compared with Solar. Our method does not rely too much on sample selection as solar. This also shows the power of the pseudo label regularization technique we proposed for imbalance PLL. 2) Consistency regularization and Mixup or Not. we ablate the contributions of two components in representation enhancement: mixup augmentation training and consistency regularization. w/o MU which removing Mixup augmentation training. w/o CR means removing consistency regularization. w/o MU+CR means removing both Mixup and consistency. The results are shown in Table 2. We find that all methods can benefit from consistency regularization and mixup. But our methods and Solar benefit more.

Influence of regularization coefficient λ and M. We conduct experiments on CIFAR10-LT with $\psi = 0.5$ and $\gamma = 100$. The result are show in Table 6. We find that selecting appropriate regularization coefficient is the key to improve the performance of Long-tail PLL. And we find that with the increase of λ, the optimal M/λ also increases, where M/λ is the regularization coefficient in formula 2. In order to better understand the role of λ and M in our method, we report accuracy on three groups of classes with different sample sizes. Recall from Section 4.1 that the class indices are sorted based on the sample size, in descending order. We divide the classes into three groups: many (1,..,[c/3]), few ([2c/3]+1,..., c) and medium (rest) shots. When M is small, as M increases, the penalty for the pseudo label of head (Many) classes increases, which in turn leads to a decrease in the accuracy of head classes. For Medium and Few classes, the accuracy will be promoted. When M is large, as M increases, the classification accuracy of all groups will decrease. The result of the combined force is that as M increases, the overall accuracy first increases and then decreases. When $\lambda = 1$,$M = 0$, our method can be PRODEN. Almost all results shown in Table 3 are better than 62.17, which is the accuracy of PRODEN in Table 1.

Table 5: The running time of one epoch on a NVIDIA V100 GPU.

	CIFAR10-LT	CIFAR100-LT
Solar	0.574s	0.769s
Ours	0.016s	0.023s
PRODEN	0.009s	0.011s

Results on fine-grained partial-label learning. In practice, semantically similar classes can lead to significant label ambiguity, as exemplified in Table 4. To test the limit of our method, we follow Solar [17] and evaluate on two fine-grained datasets: 1) CUB200-LT [27] dataset with 200 bird species; 2) CIFAR100-LT with hierarchical labels (CIFAR100-H-LT), where the candidate labels are generated within the same superclass . We set $\psi = 0.05, \gamma = 5$ for CUB-200-LT and $\psi = 0.5$, $\gamma = 20$ for CIFAR100-H-LT. These results clearly validate the effectiveness of our method, when the dataset presents severe label ambiguity.

Space Complexity Cost. The cost of pseudo labels regularization is $O(|B|c)$ in each batch, which is the same as PRODEN, where $|B|$ is the size of batch size, c is the num of class. While Solar is $|Q|c$, where the default $|Q|$ is set as $|q||B|$ and $q = 64$, T is the iteration number of SK algorithm with 50 as default.

Time Complexity Cost. The cost of pseudo labels regularization is $O(|B|c)$ in each batch, which is the same as PRODEN, where $|B|$ is the size of batch size, c is the num of class. While Solar is $O(T|Q|c)$, where the default $|Q|$ is set as $|q||B|$ and $q = 64$. The result are shown in Table 5. Experiments have proved that our pseudo labels calculation method is efficient. It only takes less than 2% of the whole training time.

4. CONCLUSION

In this work, we present a novel framework for the challenging imbalanced PLL problem. We propose a pseudo labels regularization method for PLL to keep pseudo labels away from the estimated class prior. In order to further improve performance, we have adopted techniques such as consistency loss, mixup, and sample selection. Comprehensive experiments show that our method improves baseline algorithms by a significant margin. In the future, we will explore the optimal regularization coefficient λ and M theoretically and adjust the weights between samples.

5. REFERENCES

[1] Eyal Beigman and Beata Beigman Klebanov, "Learning with annotation noise," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009, pp. 280–287.

[2] Simon C Warby, Sabrina L Wendt, Peter Welinder, Emil GS Munk, Oscar Carrillo, Helge BD Sorensen, Poul Jennum, Paul E Peppard, Pietro Perona, and Emmanuel Mignot, "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods," *Nature methods*, vol. 11, no. 4, pp. 385–392, 2014.

[3] Jia Wan and Antoni Chan, "Modeling noisy annotations for crowd counting," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3386–3396, 2020.

[4] Rong Jin and Zoubin Ghahramani, "Learning with multiple labels," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, 2002, pp. 921–928.

[5] Eyke Hüllermeier and Jürgen Beringer, "Learning from ambiguously labeled examples," *Intelligent Data Analysis*, vol. 10, no. 5, pp. 419–439, 2006.

[6] Timothee Cour, Ben Sapp, and Ben Taskar, "Learning from partial labels," *Journal of Machine Learning Research*, vol. 12, pp. 1501–1536, 2011.

[7] Timothee Cour, Benjamin Sapp, Chris Jordan, and Ben Taskar, "Learning from ambiguously labeled images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 919–926.

[8] Fei Yu and Min-Ling Zhang, "Maximum margin partial label learning," in *Proceedings of the Asian Conference on Machine Learning*, 2016, pp. 96–111.

[9] Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama, "Progressive identification of true labels for partial-label learning," in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 6500–6510.

[10] Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama, "Learning with multiple complementary labels," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2020, pp. 3072–3081.

[11] Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin, "Leveraged weighted loss for partial label learning," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2021, pp. 11091–11100.

[12] Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao, "Pico: Contrastive label disambiguation for partial label learning," in *Proceedings of the International Conference on Learning Representations*, 2022, pp. 1–18.

[13] Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao, "Pico+: Contrastive label disambiguation for robust partial label learning," *arXiv preprint arXiv:2201.08984*, 2022.

[14] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng, "Deep long-tailed learning: A survey," *arXiv preprint arXiv:2110.04596*, 2021.

[15] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar, "Long-tail learning via logit adjustment," in *International Conference on Learning Representations*. 2021, OpenReview.net.

[16] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 14567–14579, 2020.

[17] Haobo Wang, Mingxuan Xia, Yixuan Li, Yuren Mao, Lei Feng, Gang Chen, and Junbo Zhao, "Solar: Sinkhorn label refinery for imbalanced partial-label learning," in *Advances in Neural Information Processing Systems*, 2022.

[18] Marco Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.

[19] Feng Hong, Jiangchao Yao, Zhihan Zhou, Ya Zhang, and Yanfeng Wang, "Long-tailed partial label learning via dynamic rebalancing," in *The Eleventh International Conference on Learning Representations*, 2022.

[20] Dong-Dong Wu, Deng-Bao Wang, and Min-Ling Zhang, "Revisiting consistency regularization for deep partial label learning," in *Proceedings of the International Conference on Machine Learning*, 2022, pp. 24212–24225.

[21] Zheng Lian, Mingyu Xu, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao, "Arnet: Automatic refinement network for noisy partial label learning," *arXiv preprint arXiv:2211.04774*, 2022.

[22] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018, pp. 1–13.

[23] Mingyu Xu, Zheng Lian, Lei Feng, Bin Liu, and Jianhua Tao, "Dali: Dynamically adjusted label importance for noisy partial label learning," *arXiv preprint arXiv:2301.12077*, 2023.

[24] Xin Wang, Yudong Chen, and Wenwu Zhu, "A survey on curriculum learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.

[25] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, vol. 32, 2019.

[26] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang, "Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10857–10866.

[27] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona, "Caltech-ucsd birds 200," Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

[28] Ning Xu, Congyu Qiao, Xin Geng, and Min-Ling Zhang, "Instance-dependent partial label learning," in *Proceedings of the Advances in Neural Information Processing Systems*, 2021, pp. 27119–27130.

[29] Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama, "Provably consistent partial-label learning," *Advances in neural information processing systems*, vol. 33, pp. 10948–10960, 2020.