Rethinking Guidance Information to Utilize Unlabeled Samples: A Label Encoding Perspective

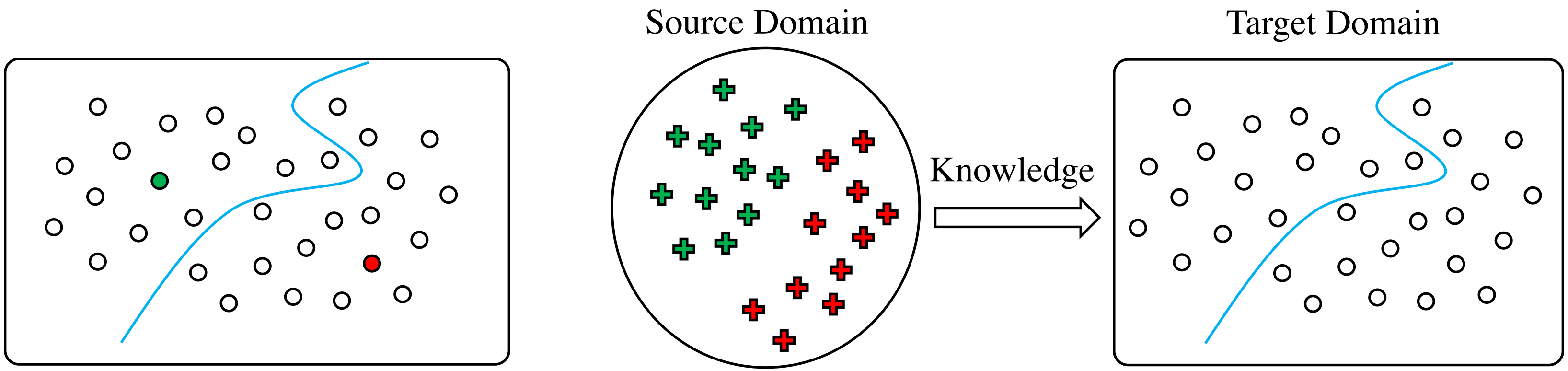
Yulong Zhang *, **Yuan Yao *** , Shuhao Chen , Pengrong Jin , Yu Zhang, Jian Jin, Jiangang Lu

2024.06

Teleinfo

Problem

How to effectively utilize unlabeled samples to handle several label insufficient scenarios?



Semi-supervised Learning

Domain Adaptation

Label insufficient scenarios

Background

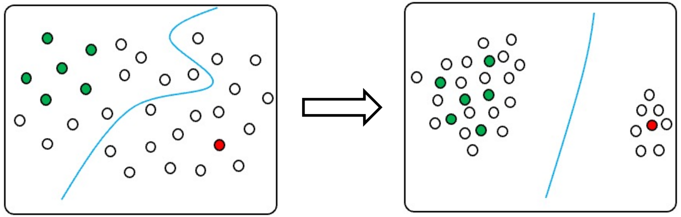
Empirical Risk Minimization (ERM), which adopts the **ground-truth label encodings** of labeled samples to guide their learning. ERM is formulated as

$$\min_{f,g} = \frac{1}{n_l} \sum_{i=1}^{n_l} \mathcal{L}\left[f(g(\mathbf{x}_i^l)), \mathbf{y}_i^l\right]$$ Ground-truth label Encoding: $[1, 0, 0]$

A vanilla extension of ERM to unlabeled samples is **Entropy Minimization (EntMin)**, which utilizes the **soft-label encodings** of unlabeled samples to guide their learning. EntMin is formulated as

$$\min_{f,g} = -\frac{1}{n_u} \sum_{i=1}^{n_u} (\tilde{\mathbf{y}}_i^u)^\top \ln \tilde{\mathbf{y}}_i^u \quad \tilde{\mathbf{y}}_i^u = f(g(\mathbf{x}_i^u)) \in \mathbb{R}^C$$ Soft-label Encoding: $[0.1, 0.7, 0.2]$

However, EntMin **emphasizes prediction discriminability while neglecting prediction diversity** [1].



For unlabeled samples, is there more precise guidance information available???

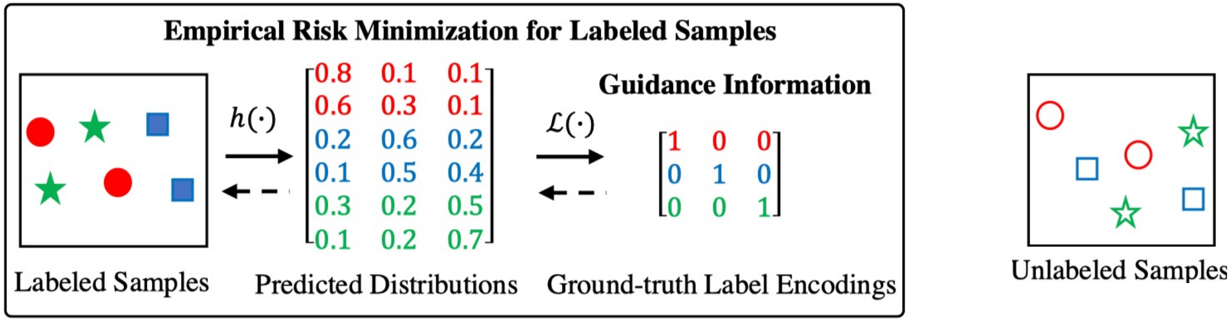
[1] Cui, S., Wang, S., Zhuo, J., Li, L., Huang, Q., and Tian, Q. Towards discriminability and diversity: Batch nuclearnorm maximization under label insufficient situations. CVPR 2020

Motivation

By analyzing the **ERM's learning objective**, we find that:

- The guidance information of the labeled samples in a specific category is the corresponding label encoding.
- There is a one-to-one correspondence between label encoding and category.

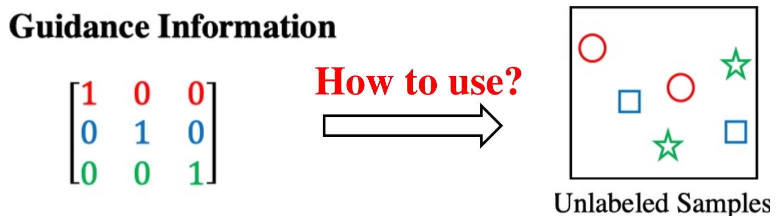
Accordingly, those label encodings remain available for unlabeled samples !!!



(a) ERM

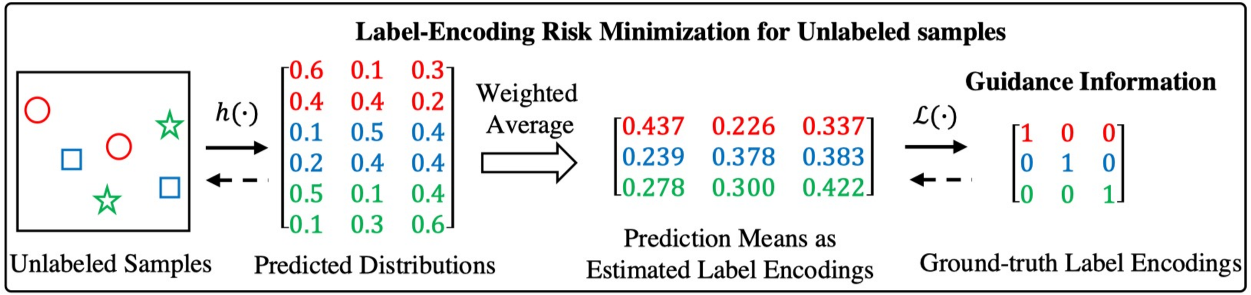
Motivation

How to utilize the label encodings to supervise the learning of unlabeled samples?



Using unlabeled samples to estimate label encoding of each category !

Using the **predicted category distribution** of unlabeled samples to **estimate label encodings** in all categories.



(b) LERM

$$\frac{[0.6\ 0.4\ 0.1\ 0.2\ 0.5\ 0.1] * \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.5 & 0.1 & 0.4 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}}{0.6 + 0.4 + 0.1 + 0.2 + 0.5 + 0.1}$$

$$= [0.437\ 0.226\ 0.337]$$

Methodology

The prediction mean for category c is defined as

$$\mathbf{m}_c^u = \frac{1}{\sum_{i=1}^{n_u} \widetilde{y}_{i,c}^u} \left(\sum_{i=1}^{n_u} \widetilde{y}_{i,c}^u \widetilde{\mathbf{y}}_i^u \right)$$

$$\frac{[0.6\ 0.4\ 0.1\ 0.2\ 0.5\ 0.1] * \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.5 & 0.1 & 0.4 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}}{0.6 + 0.4 + 0.1 + 0.2 + 0.5 + 0.1}$$

$$= [0.437\ 0.226\ 0.337]$$

Theorem 4.1. $\mathbf{m}_c^u$ *satisfies the following properties:*

(1) $\mathbf{1}^T \mathbf{m}_c^u = 1$, *where* $\mathbf{1} \in \mathbb{R}^C$ *denotes an all-ones vector.*

(2) $0 \le m_{c,j}^u \le 1, \forall j \in \{1, \dots, C\}$, *where* $m_{c,j}^u$ *denotes the* j*-th element of* $\mathbf{m}_c^u$.

(3) If $\widetilde{\mathbf{y}}_i^u$ *equals the label encoding of the ground-truth label of sample* $\mathbf{x}_i^u$ *for each* $i \in \{1, \dots, n_u\}$, *then* $\mathbf{m}_c^u$ *equals* $\mathbf{e}_c$. *Here,* $\mathbf{e}_c$ *denotes the one-hot label encoding of category* c *with its* c*-th element as 1 and other elements as 0.*

(4) If $\mathbf{m}_c^u$ *equals* $\mathbf{e}_c$ *for some* $c \in \{1, \cdots, C\}$, *then for any* $i \in \{1, \cdots, n_u\}$, $\widetilde{\mathbf{y}}_i^u$ *either equals* $\mathbf{e}_c$ *or satisfies the condition that* $\widetilde{y}_{i,c}^u = 0, 0 \le \widetilde{y}_{i,k}^u \le 1, \forall k \ne c$.

(5) If $\mathbf{m}_c^u$ *equals* $\mathbf{e}_c$ *for any* $c \in \{1, \cdots, C\}$, *then for any* $i \in \{1, \cdots, n_u\}$, $\widetilde{\mathbf{y}}_i^u$ *is a one-hot vector with only one element equal to 1 and other elements being 0.*

Based on property (3) in Theorem 4.1, we find that $\mathrm{m}_\mathrm{c}^\mathrm{u}$ could be regarded as an estimation for e_c. Accordingly, we formulate the LERM as

$$\min_{f,g} = \frac{1}{C} \sum_{c=1}^{C} \mathcal{L}(\mathbf{m}_c^u, \mathbf{e}_c)$$

LERM can ensure the prediction discriminability and diversity to some extent.

Discussion

1. Connection between LERM and ERM

Theorem 4.2. *Under the setting of supervised learning, if both the label-encoding and empirical risks utilize the same loss function which is convex w.r.t. the first input argument and* $\frac{1}{n_l}\sum_{c=1}^{C} n_c^l\, \mathcal{L}(\mathbf{m}_c^l, \boldsymbol{e}_c) \geq \frac{1}{C}\sum_{c=1}^{C} \mathcal{L}(\mathbf{m}_c^l, \boldsymbol{e}_c)$ *holds, then the label-encoding risk is upper-bounded by the empirical risk.*

2. Connection between LERM and EntMin

Theorem 4.3. *If the label-encoding risk utilizes the cross-entropy loss function, i.e.,* $\mathcal{L}(\mathbf{m}_c^u, \boldsymbol{e}_c) = -\boldsymbol{e}_c{}^T \ln \mathbf{m}_c^u$ *and the inequality* $\frac{1}{n_u}\sum_{c=1}^{C}(\sum_{j=1}^{n_u} \tilde{y}_{j,c}^u)\, \mathcal{L}(\mathbf{m}_c^u, \boldsymbol{e}_c) \geq \frac{1}{C}\sum_{c=1}^{C} \mathcal{L}(\mathbf{m}_c^u, \boldsymbol{e}_c)$ *holds, then the label-encoding risk is upper-bounded by the entropy regularization used in the EntMin.*

Application

Teleinfo

1. Semi-Supervised Learning (SSL)

$$\min_{f,g} \frac{1}{n_l}\sum_{i=1}^{n_l}\mathcal{L}_{ce}\Big[f(g(\psi(\mathbf{x}_i^l))), \mathbf{y}_i^l\Big] + \frac{\mu}{n_l}\sum_{i=1}^{n_l}\mathcal{L}_{ce}\Big[f(g(\Psi(\mathbf{x}_i^l))), \mathbf{y}_i^l\Big] + \alpha\mathcal{L}_{ssl} + \frac{\lambda}{C}\sum_{c=1}^{C}\Big[\mathcal{L}(\mathbf{w}_c^u, \mathbf{e}_c) + \mu\mathcal{L}(\mathbf{s}_c^u, \mathbf{e}_c)\Big]$$

2. Unsupervised Domain Adaptation (UDA)

$$\min_{f,g} \frac{1}{n_s}\sum_{i=1}^{n_s}\mathcal{L}_{ce}\Big[f(g(\mathbf{x}_i^s)), \mathbf{y}_i^s\Big] + \alpha\mathcal{L}_{uda} + \frac{\lambda}{C}\sum_{c=1}^{C}\mathcal{L}(\mathbf{m}_c^u, \mathbf{e}_c)$$

3. Semi-supervised Heterogeneous Domain Adaptation (SHDA)

$$\min_{f,g_s,g_t} \frac{1}{n_s}\sum_{i=1}^{n_s}\mathcal{L}_{ce}\Big[f(g_s(\mathbf{x}_i^s)), \mathbf{y}_i^s\Big] + \frac{1}{n_l}\sum_{i=1}^{n_l}\mathcal{L}_{ce}\Big[f(g_t(\mathbf{x}_i^l)), \mathbf{y}_i^l\Big] + \alpha\mathcal{L}_{shda} + \frac{\lambda}{C}\sum_{c=1}^{C}\mathcal{L}(\dot{\mathbf{m}}_c^u, \mathbf{e}_c) + \tau(\|f\|^2 + \|g_s\|^2 + \|g_t\|^2)$$

Experiments: Evaluation on SSL Tasks

Teleinfo

Table 1. Accuracy (%) comparison on the CIFAR-10, CIFAR-100, DTD, and ImageNet-1K datasets under the SSL setting. The best performance of each task is marked in bold and the best performance in each comparison group is underlined.

Dataset	CIFAR-10			CIFAR-100			DTD			ImageNet-1K	
# Label per category	1		4	1		4	1		4	100	
	Top-1	Top-5	Top-1	Top-1	Top-5	Top-1	Top-1	Top-5	Top-1	Top-1	Top-5
ERM	32.24	78.16	57.04	23.58	47.51	47.18	31.22	58.99	50.66	44.98	69.00
ERM + EntMin	28.17	71.05	59.62	15.32	43.95	45.40	21.55	51.65	50.96	49.26	72.60
ERM + BNM	27.02	70.37	52.46	21.79	47.72	58.90	28.55	54.61	48.26	49.81	72.73
ERM + LERM	<u>38.22</u>	<u>80.82</u>	<u>75.57</u>	<u>30.15</u>	<u>61.33</u>	<u>60.19</u>	<u>34.84</u>	<u>63.51</u>	<u>53.14</u>	<u>50.83</u>	<u>74.11</u>
FlexMatch	40.86	84.75	86.66	16.49	42.40	65.11	33.39	58.48	54.96	50.34	75.02
FlexMatch + EntMin	43.79	87.69	86.56	13.00	42.83	67.32	32.20	58.49	54.91	53.26	76.99
FlexMatch + BNM	41.95	78.73	86.57	15.04	43.54	64.46	31.31	57.31	55.04	55.12	78.62
FlexMatch + LERM	<u>53.69</u>	<u>89.18</u>	<u>88.28</u>	<u>19.50</u>	<u>46.00</u>	**<u>69.65</u>**	<u>34.42</u>	<u>58.51</u>	<u>55.11</u>	**<u>56.69</u>**	**<u>79.79</u>**
DST	51.11	91.76	88.05	32.92	64.65	66.80	34.88	61.99	56.40	50.34	75.94
DST + EntMin	45.46	92.41	87.85	25.48	60.92	66.79	32.32	62.27	56.13	53.82	76.28
DST + BNM	55.03	91.75	88.49	32.15	65.16	67.27	36.08	64.06	56.51	54.28	76.56
DST + LERM	**<u>62.04</u>**	**<u>93.09</u>**	**<u>89.71</u>**	**<u>43.78</u>**	**<u>70.37</u>**	<u>68.65</u>	**<u>38.19</u>**	**<u>67.39</u>**	**<u>57.45</u>**	<u>54.60</u>	<u>76.87</u>

Experiments: Prediction Discriminability Analysis

Teleinfo

We can observe that ERM+EntMin and ERM+LERM obtain much lower entropy values than ERM. Those results show that both EntMin and LERM achieve good prediction discriminability.

$$\min_{f,g} = \frac{1}{C}\sum_{c=1}^{C} \mathcal{L}(\mathbf{m}_c^u; \mathbf{e}_c)$$

Table 7. Prediction discriminability comparison on the CIFAR-10 dataset under the SSL setting.

Method	Entropy
ERM	0.3832
ERM + EntMin	0.0266
ERM + LERM	0.0440

Experiments: Prediction Diversity Analysis

Teleinfo

We rebuild the SSL task on the CIFAR-10 dataset into a category-imbalanced setting. We can see that compared with ERM + EntMin, ERM + LERM is less susceptible to the impact of category imbalance. Those results indicate that the LERM can effectively preserve prediction diversity even in category-imbalanced scenarios.

$$\min_{f,g} = \frac{1}{C}\sum_{c=1}^{C} \mathcal{L}(\mathbf{m}_c^u, \mathbf{e}_c)$$

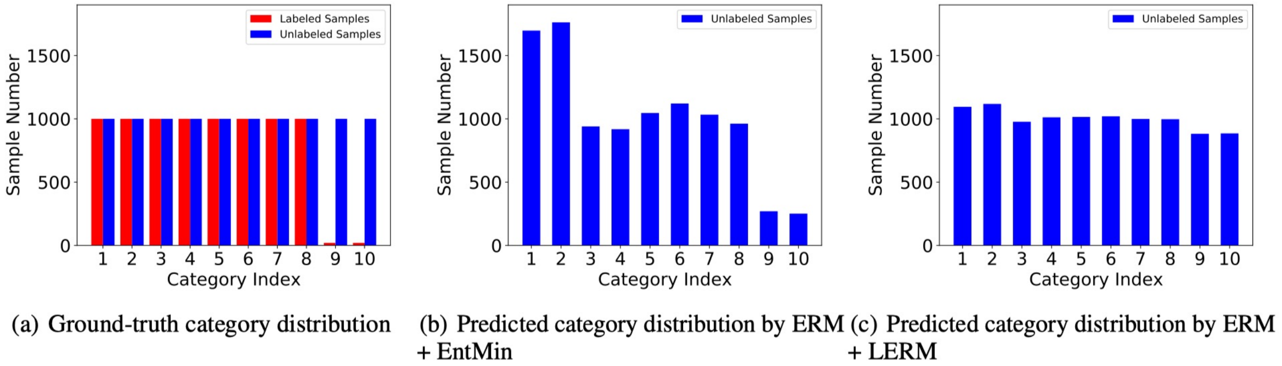


Figure 3. Empirical evaluation of prediction diversity on the SSL task on CIFAR-10 dataset under the class-imbalanced setting. (a) The ground-truth category distributions of the labeled and unlabeled samples. (b) The predicted category distribution of the unlabeled samples by ERM + EntMin. (c) The predicted category distribution of the unlabeled samples by ERM + LERM.

Thank you all for your time and participation!

Paper: https://arxiv.org/abs/2406.02862

Code: https://github.com/zhangyl660/LERM

Contact: yaoyuan.hitsz@gmail.com