

Learning from Multiple Unlabeled Datasets in Weakly Supervised Multi-Class Classification: A Class-Prior Based Regularization Approach

Yuting Tang 47206094

Department of Complexity Science and Engineering
Graduate School of Frontier Science, The University of Tokyo

I. INTRODUCTION

Nowadays, supervised classification has achieved great success. However, collecting massive labeled data is costly or sometimes impossible in some scenarios. Thus, it is desired for machine learning techniques to work with weakly-labeled data, which can reduce labeling costs and still reach a good result.

Learning from multiple unlabeled datasets is a weakly supervised learning setting where the supervision information is the class-prior of each unlabeled dataset rather than labeled data. The task is to learn an instance-level classifier from these unlabeled datasets. This setting has a wide range of applications in the real world. For example, we sometimes only know the class-prior to votes in each region during the voting process, not each vote. We hope to predict the votes of a single voter from such weakly labeled data.

In this thesis, we propose a novel class-prior based approach to deal with the multi-class classification in learning from multiple unlabeled datasets. To the best of our knowledge, it is the first time this problem setting has been discussed in multi-class classification. Firstly, we find that the Bayes optimal classifier can be obtained when the risk of considering all data in one unlabeled dataset to belong to the same class is equal to the corresponding misclassification rate, i.e., the sum of class-priors of the remaining classes within the corresponding distribution of this dataset. Following this idea, we propose a new learning objective to enforce each risk term close to their corresponding misclassification rate, which we call “risk floating”. Then we further propose to use it as a regularization of the backward corrected risk [1], an unbiased estimator of the classification risk. In this way, we can avoid the unidentifiable problem of risk floating in the finite-sample case, and the overfitting problem that is easy to occur in backward correction [2], [3]. Finally, we validate the effectiveness of our proposed method through experiments.

II. PROBLEM

In this section, we formalize the problem of learning from multiple unlabeled datasets. We consider the multi-class classification problem in learning from multiple unlabeled datasets. Suppose M is the number of classes, and N is the number of unlabeled datasets. We define $\Delta_M = \{\gamma \in \mathbb{R}_+^M : \sum_{m=1}^M \gamma_m = 1\}$ to be a probability simplex, where \mathbb{R}_+^M is the non-negative real coordinate space of dimension M and γ_m is the m -th element of γ . Let $\theta_n = (\theta_{n,1}, \dots, \theta_{n,M})^\top \in \Delta_M$, $n = 1, \dots, N$ be the class-prior of M classes in distribution \mathcal{D}'_n . Supposing $\mathcal{D}_m := p(\mathbf{x}|Y = m)$ is the class-conditional distribution of class m , the dataset \mathcal{X}_n is generated as follow:

$$\mathcal{X}_n \sim \mathcal{D}'_n = \sum_{m=1}^M \theta_{n,m} \mathcal{D}_m.$$

We hope to train an instance-level classifier on unlabeled datasets \mathcal{X}_n , $n = 1, \dots, N$ and get low classification error on the test dataset $\mathcal{X}_{\text{test}} \sim \sum_{m=1}^M \pi_m \mathcal{D}_m$, where π_m , $m = 1, \dots, M$ are class priors of M classes in the test dataset.

III. PROPOSED METHOD

In this section, we propose a simple but effective method to solve the problem of learning from multiple unlabeled datasets. We propose a risk floating method and extend the backward correction method [1] to the problem of learning from multiple unlabeled datasets. The idea is to use the learning objective of risk floating as a regularization of the backward corrected risk.

A. Risk Floating

Here, we first prove a lemma that characterizes the condition a good classifier needs to satisfy, then propose a learning objection based on this lemma. To introduce this lemma, we define two partial risks: $R_{n,(m)}(\mathbf{g}; \ell_{01}) := \mathbb{E}_{(\mathbf{x}, Y) \sim \mathcal{D}'_n} [\ell_{01}(\mathbf{g}(\mathbf{x}), m)]$ and $R_{(m)}^t(\mathbf{g}; \ell_{01}) := \mathbb{E}_{(\mathbf{x}, Y) \sim \mathcal{D}_m} [\ell_{01}(\mathbf{g}(\mathbf{x}), t)]$. The brief explanation is $R_{n,(m)}(\mathbf{g}; \ell_{01})$ represents the risk of considering all samples generated from mixture distribution \mathcal{D}'_n belonging to class m , and $R_{(m)}^t(\mathbf{g}; \ell_{01})$ represents the risk of considering data generated from \mathcal{D}_m to be class t . The risks $R_{n,(m)}(\mathbf{g}; \ell)$ and $R_{(m)}^t(\mathbf{g}; \ell)$ below are $R_{n,(m)}(\mathbf{g}; \ell_{01})$ and $R_{(m)}^t(\mathbf{g}; \ell_{01})$ calculated by a surrogate loss ℓ , respectively.

Lemma 1. (Float condition) Suppose we have N unlabeled datasets such that $N \geq M$. For any classifier \mathbf{g} ,

$$R_{n,(m)}(\mathbf{g}; \ell_{01}) = 1 - \theta_{n,m}, \quad (1)$$

where $\theta_n = (\theta_{n,1}, \dots, \theta_{n,M})^\top \in \Delta_M$ and $\theta_i \neq \theta_j, i \neq j$, if and only if for $t = 1, \dots, M$,

$$R_{(m)}^t(\mathbf{g}; \ell_{01}) = \begin{cases} 0 & \text{if } m = t, \\ 1 & \text{otherwise.} \end{cases} \quad (2)$$

Motivated by Lemma 1, we hope that each partial risk $R_{n,(m)}(\mathbf{g}; \ell)$ to be close to $1 - \theta_{n,m}$. Then we propose a learning objective named risk floating:

$$R_{\text{float}}(\mathbf{g}; \ell) := \sum_{n=1}^N \sum_{m=1}^M \lambda_{n,m} |R_{n,(m)}(\mathbf{g}; \ell) - (1 - \theta_{n,m})|, \quad (3)$$

where $\lambda_{n,m}$ s are non-negative hyperparameters.

B. Backward corrected risk

Here, we extend the backward correction method to the problem of learning from multiple unlabeled datasets.

The classification risk can be written with the class-conditional distributions as

$$R(\mathbf{g}; \ell) = \sum_{m=1}^M \pi_m \mathbb{E}_{X \sim \mathcal{D}_m} [\ell(\mathbf{g}(X), Y)], \quad (4)$$

where $\pi_m := p(Y = m)$ is the class-prior probability. Since we cannot get the class label Y , we modify the loss function to get the backward corrected form of the classification risk. It can be written as

$$R_{\text{U-BC}}(\mathbf{g}; \ell) = \sum_{n=1}^N \sum_{m=1}^M (\theta_{n,m} k_{n,1} \mathbb{E}_{(X,Y) \sim \mathcal{D}_m} [\ell(\mathbf{g}(X), 1)] + \dots + \theta_{n,m} k_{n,M} \mathbb{E}_{(X,Y) \sim \mathcal{D}_m} [\ell(\mathbf{g}(X), M)]). \quad (5)$$

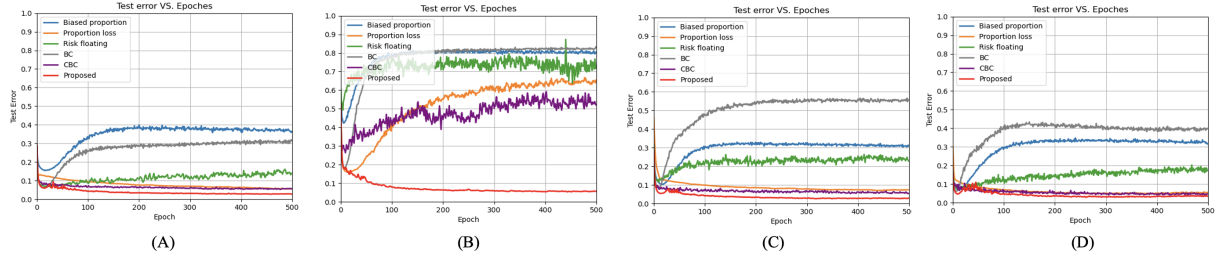


Fig. 1: Comparing with baseline methods. (A) and (B) shows the results with symmetrical class-prior matrix such that $a = 0.5, b = 0.05$ and $a = 0.1, b = 0.09$, respectively. (C) shows the result on asymmetrical class-prior matrix. The result of learning from 20 unlabeled datasets is presented in (D).

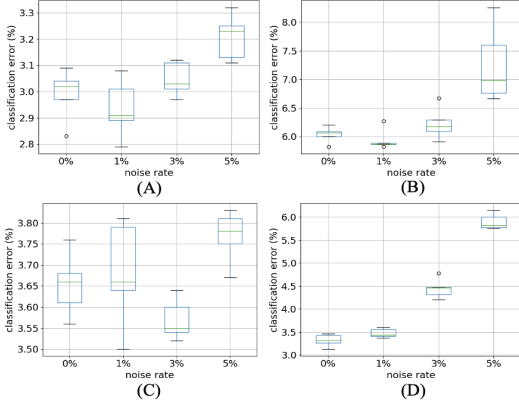


Fig. 2: Robustness against inaccurate class priors. (A), (B), (C), and (D) have the same meaning as Fig. 1.

By comparing (4) and (5), the constant coefficients $k_{n,m}, n = 1, \dots, N, m = 1, \dots, M$ can be specified, then the classification risk is calculable.

C. Proposed learning objective

Finally, we give our learning objective. We use the learning objective of risk floating as a regularization term of the backward corrected risk:

$$R(\mathbf{g}; \ell) = R_{\text{U-BC}}(\mathbf{g}; \ell) + \alpha R_{\text{float}}(\mathbf{g}; \ell), \quad (6)$$

where $\alpha \geq 0$ is added to balance between these two risk functions. We determined the hyperparameters $\lambda_{n,m}$ in the risk floating part as $\lambda_{n,m} = |k_{n,m}|$. This choice means that the term that receives more attention in the backward corrected risk also receives more attention in risk floating. Since $\lambda_{n,m}$ are all non-negative, we applied the absolute value operator.

IV. EXPERIMENTS

In this section, we show the experiments conducted on the MNIST dataset. The base model was a 5-layer MLP, the optimizer was Adam, and the learning rate was $1e-4$. The cross-entropy loss was used as a surrogate loss. For the data generation process, if the class-prior of an unlabeled dataset was θ and dataset size was S , $\theta_m \times S$ samples would be randomly chosen from class m . All methods used the same architecture, batch size, and training procedure. We call Θ as the class-prior matrix such that $[\Theta]_{n,m} = \theta_{n,m}$.

The proposed method was compared with the following approaches: (i). Biased proportion: Consider the label of the largest category in each unlabeled dataset as the label of all samples in the dataset. Then perform supervised learning from such pseudo label samples. (ii). Proportion Loss [4]: Use class priors of each unlabeled dataset as weak supervision. (iii). Risk floating method: Treat $\hat{R}_{\text{float}}(\mathbf{g})$ as a learning objective. (iv). Backward

correction (BC) method: Treat $\hat{R}_{\text{BC}}(\mathbf{g})$ as a learning objective. (v). Corrected backward correction (CBC) method: A consistently corrected risk estimator of $\hat{R}_{\text{BC}}(\mathbf{g})$ inspired by [3].

A. Comparing with baseline methods

The first two experiments were performed on ten unlabeled datasets with class priors

$$\theta_{n,m} = \begin{cases} a + b & n = m, \\ b & n \neq m, \end{cases}$$

where a and b are constants satisfying $a > b$ and $a + 10b = 1$. We call this kind of class-prior matrix Θ a symmetric class-prior matrix, because it can be written as a symmetric matrix. Figs. 1 (A) and (B) show the test errors of them. Another experiment was performed on ten unlabeled datasets with the asymmetric class-prior matrix. Here, Θ is a diagonal-dominated matrix whose values on the diagonals are larger than the other values, and they can be different from each other. The result is shown in Fig. 1 (C). We also conducted experiments on more than ten unlabeled datasets. Compared with the previous experiments, this is more in line with the real-world situation because the number of unlabeled datasets is usually higher than the number of classes. The result of learning from 20 unlabeled datasets with different class priors is shown in Fig. 1 (D). The experimental results show that the proposed method is better than the other methods in all cases.

B. Robustness against Inaccurate Class Priors

We also designed experiments that add random noise to all class priors to simulate real-world situations where class priors may be noisy. In the experiment, we added different levels (1%, 3%, and 5%) of perturbation to each term of Θ to obtain Θ' , so that the method would treat noisy Θ' as the true Θ during the whole learning process. From Fig. 2, we know that the proposed method is rarely affected by the noisy class priors and can be safely applied in noisy problems.

V. CONCLUSION

In this thesis, we proposed a simple but effective method to solve the problem of learning from multiple unlabeled datasets. We demonstrated the effectiveness of the proposed method through experiments.

REFERENCES

- [1] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [2] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," in *NIPS*, 2017.
- [3] N. Lu, T. Zhang, G. Niu, and M. Sugiyama, "Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1115–1125.
- [4] F. X. Yu, K. Choromanski, S. Kumar, T. Jebara, and S.-F. Chang, "On learning from label proportions," *arXiv:2007.01807*, 2015.