FY2023/2023年度

Imperfect Information Learning Team

Masashi Sugiyama

不完全情報学習チーム　杉山 将

RIKEN Center for Advanced Intelligence Project

Our Vision and Social Impact:

- Develop **trustworthy machine learning** methods/algorithms that can cope with imperfect training information like **distribution shift**, noisy labels, partial labels, and **pseudo-supervision**.
- Enable machine learning for real-world applications in imperfect or adversarial deployment environments such as robust image/video classification and sample-/label-efficient text classification.

Members

- Masashi Sugiyama (Team Leader)
- Takashi Ishida (Research Scientist)
- Masahiro Fujisawa (SPDR)
- Zhen-Yu Zhang (Postdoc)
- Gang Niu (Senior Research Scientist)
- Shuo Chen (Research Scientist)
- Guillaume Braun (Postdoc)
- Okan Koc (Postdoc)

Learning under Distribution Shift

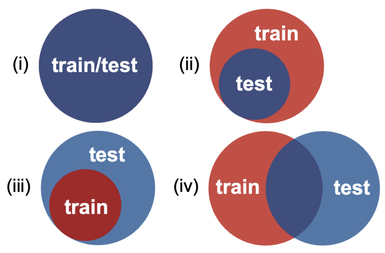
Distribution shift is quite common nowadays. Supervised learning requires the training and test data to be identically distributed. In practice, however, it may be infeasible to collect sufficient test-distributed training data. Learning under distribution shift is aimed at training a classifier using non-test-distributed data, where the classifier is expected to make predictions for future test data as if it is trained using test-distributed data.

- Traditionally, solving distribution shift problems requires that the training distribution must be "bigger" or "wider" than the test distribution. This is an open issue since 1949. We addressed the open issue and removed the requirement (Fang et al., NeurIPS 2023).
- Continuous distribution shift problems require appropriate reuse of historical data. We studied the continuous covariate shift case and proposed an online ensemble method that approximates the optimal model sequence at any time (Zhang et al., NeurIPS 2023).

Generalized Importance Weighting

- **Distribution Shift (DS):** There are two levels of DS:
  1. Only the data distribution changes, while its support (the set where the probability density is non-zero) remains the same;
  2. The support of the data distribution also changes.

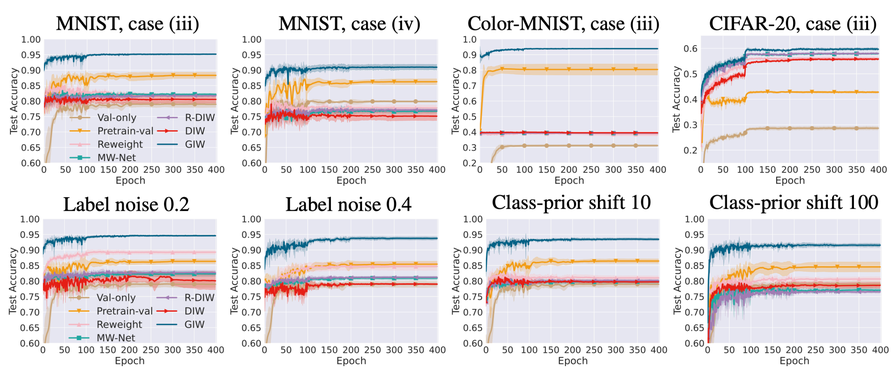
  There are four cases when considering the relationship between the training and test support:



- **Importance Weighting (IW):** IW has been a golden solver for DS problems. IW works as follows: first, estimate a test-over-training density ratio function to obtain importance weights; then, train a classifier with the importance weights by weighted classification.
- **Motivation:** IW requires the support of the test distribution to be included in the support of the training distribution, i.e., case (i) or (ii). However, this assumption may not be satisfied in practice. Indeed, cases (iii) and (iv) are more common nowadays but still under-explored.
- **Methodology:** We proved that the objective of IW is good in cases (i) and (ii) and bad in cases (iii) and (iv). Then, we proposed generalized IW (GIW) as a universal solver for DS problems: GIW could handle cases (iii) and (iv) and would reduce to IW in cases (i) and (ii). We also designed practical implementations of GIW based on the one-class support vector machine. We finally demonstrated the effectiveness of GIW by extensive experiments on popular benchmark datasets such as MNIST and CIFAR-20.

Results on support shift (DS solely comes from the support mismatch)

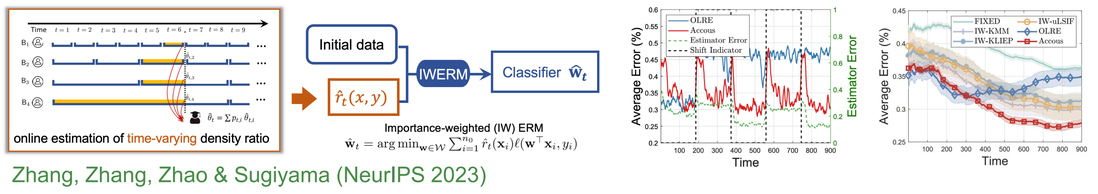
Results on distribution-and-support shift



Fang, Lu, Niu & Sugiyama (NeurIPS 2023 spotlight)

Adapting to Continuous Covariate Shift

- **Continuous Covariate Shift:** The test distribution can continuously evolve over time. Hence, there is a single training distribution but there are many test distributions at different time.
- **Motivation:** Most existing covariate-shift approaches focus on static covariate shift problems but struggle to handle continuous covariate shift in the entire sequential prediction process.
- **Methodology:** We proposed to handle continuous covariate shift using adaptive importance weighting, where the weights are learned via online density-ratio estimation. We developed an online ensemble learning method to handle the unknown non-stationary environments. It works as follows: a set of base models that exploit different lengths of historical test data, and a meta model that combines them to make the final prediction.



Zhang, Zhang, Zhao & Sugiyama (NeurIPS 2023)

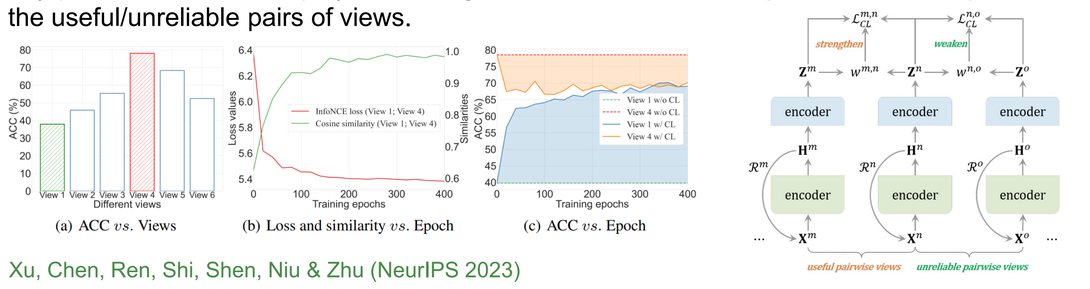
Contrastive Learning

Contrastive learning is quite popular nowadays. Supervised learning requires that every training instance has a label as the model output target. In practice, however, it may be difficult to obtain sufficient labeled data. Contrastive learning is aimed at training a good task-independent representation by pairwise similarity/dissimilarity pseudo-supervision. Then, it is more sample-/label-efficient to train a simpler task-specific decision model.

- Multi-view contrastive learning may suffer from conflicting views. We proposed how to measure the view disagreement and reweight different views (Xu et al., NeurIPS 2023).
- In knowledge distillation, the teacher and the student have different training data. We proposed a contrastive network block to address the data gap (Tang et al., ICCV 2023).
- Graph contrastive learning does not consider the node importance. We proposed to emphasize certain important nodes in learned graph embeddings (Wan et al., TNNLS).

Self-Weighted Contrastive Learning among Multiple Views

- **Multi-View Contrastive Learning:** Every instance is present in multiple views; the algorithm learns view encoders (one for each view) by aligning different views of the same instance.
- **Motivation:** The multiple views are not necessarily consistent --- quality difference can often be observed, and different views may even semantically conflict with each other.
- **Methodology:** We proposed a self-weighted loss that measures the disagreement between any pair of views and employs the disagreement, so that we can put more/less emphasis on the useful/unreliable pairs of views.

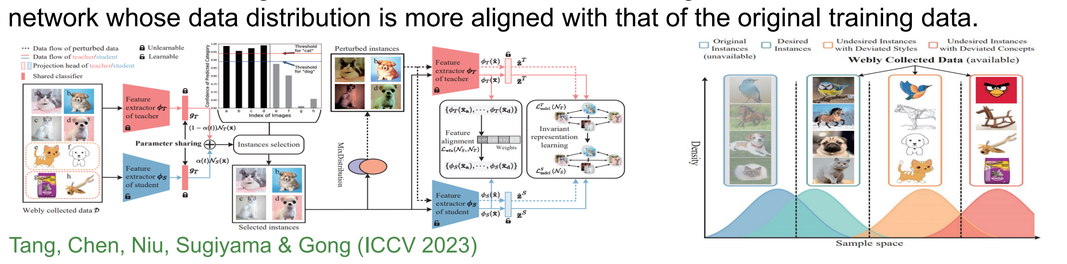


(a) ACC $vs.$ Views (b) Loss and similarity $vs.$ Epoch (c) ACC $vs.$ Epoch

Xu, Chen, Ren, Shi, Shen, Niu & Zhu (NeurIPS 2023)

Contrastive Learning based Data-Free Knowledge Distillation

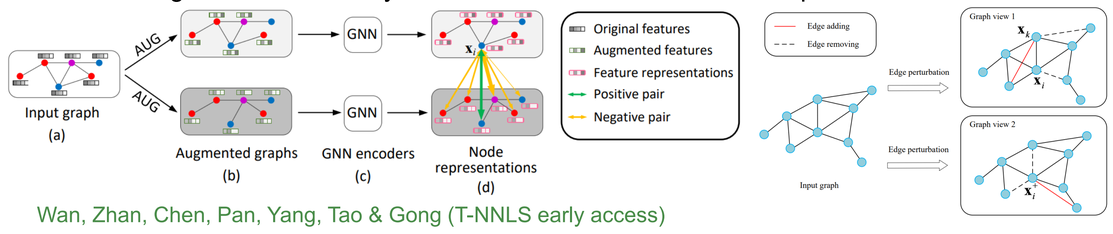
- **Data-Free Knowledge Distillation:** The training data for the teacher network are no longer available; as a result, the training data for the student network are from the Internet.
- **Motivation:** The existing methods paid no attention to distribution shift between the original training data (for the teacher) and the new webly collected data (for the student).
- **Methodology:** We proposed to dynamically select useful training data among the new webly collected data according to the weighted mean scores of the teacher and student networks. Furthermore, we designed a contrastive network block to generate new data for the student network whose data distribution is more aligned with that of the original training data.



Tang, Chen, Niu, Sugiyama & Gong (ICCV 2023)

Boosting Graph Contrastive Learning via Adaptive Sampling

- **Graph Contrastive Learning:** Every training data is given as a graph; the algorithm learns graph embeddings (one for each graph) by pushing away (negative) pairs of graph nodes.
- **Motivation:** The nodes used for constructing negative pairs are uniformly sampled from the augmented views, but some of them may be semantically positive (i.e., false negatives).
- **Methodology:** We proposed to adaptively encode the node importance and encourage only utilizing the most important graph nodes (indeed, most confident/informative negative pairs). We also proposed an auxiliary polarization regularization to suppress the negative impact of the false negatives and thereby further enhance the discriminative power.



Wan, Zhan, Chen, Pan, Yang, Tao & Gong (T-NNLS early access)

References

- T. Fang, N. Lu, G. Niu, and M. Sugiyama. "Generalizing importance weighting to a universal solver for distribution shift problems", ***NeurIPS 2023* (spotlight)**.
- Y.-J. Zhang, Z.-Y. Zhang, P. Zhao, and M. Sugiyama. "Adapting to continuous covariate shift via online density ratio estimation", ***NeurIPS 2023***.
- J. Xu, S. Chen, Y. Ren, X. Shi, H.-T. Shen, G. Niu, and X. Zhu. "Self-weighted contrastive learning among multiple views for mitigating representation degeneration", ***NeurIPS 2023***.
- J. Tang, S. Chen, G. Niu, M. Sugiyama, and C. Gong. "Distribution shift matters for knowledge distillation with webly collected images", ***ICCV 2023***.
- S. Wan, Y. Zhan, S. Chen, S. Pan, J. Yang, D. Tao, and C. Gong. "Boosting graph contrastive learning via adaptive sampling", ***IEEE TNNLS*** (early access).

Other Selected Publications

- T. Ishida, I. Yamane, N. Charoenphakdee, G. Niu, and M. Sugiyama. "Is the performance of my deep network too good to be true? A direct approach to estimating the Bayes error in binary classification", ***ICLR 2023* (oral)**.
- S. Xia, J. Lv, N. Xu, G. Niu, and X. Geng. "Towards effective visual representations for partial-label learning", ***CVPR 2023***.
- R. Dong, F. Liu, H. Chi, T. Liu, M. Gong, G. Niu, M. Sugiyama, and B. Han. "Diversity-enhancing generative network for few-shot hypothesis adaptation", ***ICML 2023***.
- H. Wei, H. Zhuang, R. Xie, L. Feng, G. Niu, B. An, and Y. Li. "Mitigating memorization of noisy labels by clipping the model prediction", ***ICML 2023***.
- Z. Wei, L. Feng, B. Han, T. Liu, G. Niu, X. Zhu, and H. Shen. "A universal unbiased method for classification from aggregate observations", ***ICML 2023***.
- Z. Yan, X. Li, K. Wang, S. Chen, J. Li, and J. Yang. "Distortion and uncertainty aware loss for panoramic depth completion", ***ICML 2023***.
- J. Zhu, G. Yu, J. Yao, T. Liu, G. Niu, M. Sugiyama, and B. Han. "Diversified outlier exposure for out-of-distribution detection via informative extrapolation", ***NeurIPS 2023***.
- W. Wang, L. Feng, Y. Jiang, G. Niu, M.-L. Zhang, and M. Sugiyama. "Binary classification with confidence difference", ***NeurIPS 2023***.
- M.-K. Xie, J.-H. Xiao, H.-Z. Liu, G. Niu, M. Sugiyama, and S.-J. Huang. "Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning", ***NeurIPS 2023***.
- C. Gong, Y. Ding, B. Han, G. Niu, J. Yang, J. You, D. Tao, and M. Sugiyama. "Class-wise denoising for robust learning under label noise", ***IEEE TPAMI***, vol. 45, no. 3.
- S. Yang, S. Wu, E. Yang, B. Han, Y. Liu, M. Xu, G. Niu, and T. Liu. "A parametrical model for instance-dependent label noise", ***IEEE TPAMI***, vol. 45, no. 12.