Investigating Active Positive-Unlabeled Learning with Deep Networks

Kun Han[1], Weitong Chen[1] and Miao Xu[1,2]

University of Queensland, Australia[1]
RIKEN Japan 103−0027[2]

Motivation

In the real world, some classification problems are based on only positive data and unlabeled data available, which is recognized as PU learning. One significant factor of the model's performance for PU learning is the number and the quality of positive data. However, labelling data is always expensive, how to effectively get new positive instances that could benefit the training model is the new challenge.

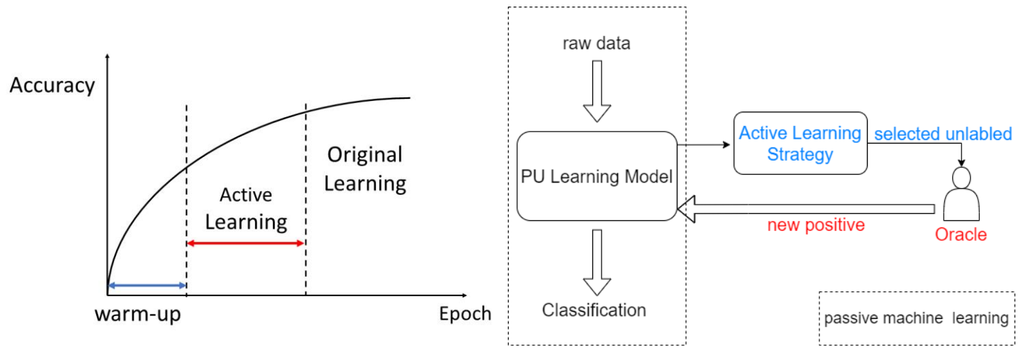
One straightforward solution is active learning, selecting the informative instances and labelling with the involvement of human beings. Inspired by this, we investigate deep active learning models based on uncertainty strategies for PU learning. The motivation in this work is:

- For previous research, active learning models for PU learning are not based on deep networks.
- The previous methods require many computation resources and are not model relevant.

Contribution

- Proposing an active learning framework based on the large-small loss trick during the training process for deep learning models.
- Investigating Various Uncertainty query strategies based on confidence, variance and density for the proposed active learning framework.

Methodology



Framework

- **Warm-up**

Start from a warm-up, which guarantees the baseline of the model's performance.

- **Active learning**
  - Calculate loss

For unlabeled data, calculate loss using ground truth negative.

  - Select instances to query

According to the query strategy, select the most informative unlabeled data to get the ground truth label of them.

  - Update dataset with new positive.
- **Continue original learning**

Algorithm

Input: Training data P and U

Parameters: query size, query epoch range

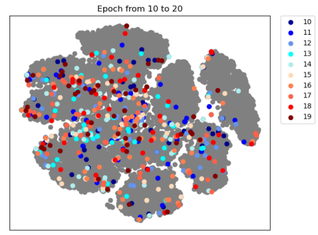
// set the parameters of nnPU
Initialization;
for *epoch* < MAX EPOCH do
  train as nnPU;
  if *epoch* within *query epoch range*
    Select and query *K* instances
    // *K* = *query size*
    Update the training P data;
  end
end

Query Strategy	Selection Criterion
Least Confidence	Loss around 0.5
Uncertainty and Density	Loss and density
Variance Max	Variances of losses
Window-sized Variance Max	Losses in range
Baseline	Loss around 1
Random Sampling	Random

Experiment

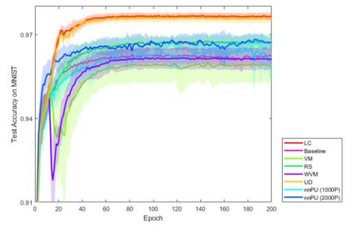
- Investigate the effect of query strategy and parameters on the MNIST dataset.
- The compared evaluation is two state-of-the-art methods, including nnPU, combined nnPU and self-PU.

Diversity

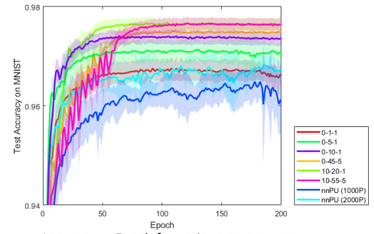


Selected instances for LC from epoch 10 to 20

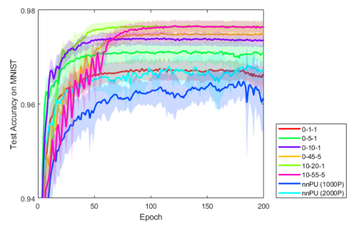
Accuracy vs. Epoch



Accuracy vs Epoch for various strategies



Accuracy vs Epoch for various parameters



Accuracy vs Epoch on MNIST for various methods

Comparison

Baseline	RS	LC	UD	VM	WVM
95.93 (0.86)	96.32 (0.98)	**97.27 (1.17)**	97.21 (1.16)	95.46 (0.95)	95.66 (1.09)

Start	0	0	0	0	0	10	10	10
Stop	1	5	10	20	45	20	30	55
Step	1	1	1	2	5	1	2	5
ACC	96.46	96.89	97.15	97.14	97.01	**97.27**	97.08	97.01
STD	0.85	0.85	0.92	1.00	1.11	**1.17**	1.17	1.25

Method	nnPU	nnPU*	Self-PU*	RS	UDALPU (LC)
Test Accuracy	95.91 ± 0.75	96.43± 0.94	95.15 ± 0.13	96.32 ± 0.98	**97.27±1.17**

Conclusion

- we investigated deep active learning methods by proposing querying strategies based on the large-small-loss trick of neural networks.
- The querying strategies are uncertainty based and require querying from the early stage of learning and also stop early before the training ends.
- Our UDALPU methods perform better in experiments.