CoConut: Co-Classification with Output Space Regularization

Sameh Khamis

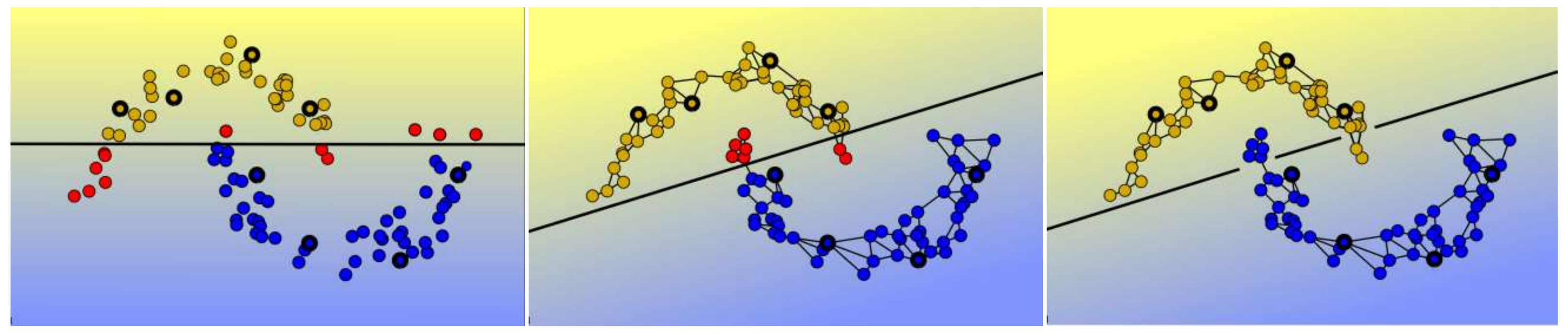
Christoph H. Lampert

Motivation

- In many real-world applications the samples to be classified occur in batches
  - Words in a document
  - Images in a photo collection
  - Stocks in a portfolio
- Can we do better at test-time by exploiting this fact and without re-training our classifiers?

Co-Classification

- We are interested in the task of jointly classifying multiple, otherwise independent, data samples
- Consider the situation of a linear classifier
  - Efficient to train
  - Generalizes well
  - Has a decision hypersurface that might not reflect the class boundaries
- Given a trained classifier and enough test samples, we can modulate its decision surface at test-time so that, for example, it does not cross high density regions



Approach

- Co-Classification with Output Space Regularization
  - Formulated as regularized risk minimization
  - Does not require classifier re-training
  - Can handle test-time additional data modalities
- Formally:

$$y^* = \operatorname*{argmin}_{y \in \mathcal{Y}^n} \; -\sum_{l=1}^{L} [\![y_i = l]\!] f_l(x_i) + \lambda \Omega(y)$$

- The regularizer Ω penalizes undesirable label combinations and λ controls its strength
- In our choice of the regularizer we encode the inductive bias we have about the problem
  - Cluster assumption

$$\Omega_S(g) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|N_i|} \sum_{x_j \in N_i} w_{ij} \delta_{ij}(g)$$

  where $\delta_{ij}(g) := [\![g(x_i) \neq g(x_j)]\!]$ indicate whether the label changed between two neighboring samples, where the neighborhood structure can be constructed:
    - ✓ With respect to the original features
    - ✓ With respect to an additional modality
    - ✓ Through a prior or side information
  - Class label distribution

$$\Omega_D(g) = \sum_{l=1}^{L} |p_l(g) - Q_l|$$

  where $p_l(g) = \frac{1}{n} \sum_{i=1}^{n} [\![g(x_i) = l]\!]$ is the label proportion that should match the target value Q_l
- The resulting labeling problem is optimized using a combination of discrete optimization and Lagrangian Relaxation

Experiments

- We evaluated CoConut on six different datasets: four image and two network datasets
- We used the features provided by the original authors
- CoConut improves over the baselines, whether through an additional data modality or a prior on class label distribution
- Unsurprisingly, with more data at test-time the improvement is generally more consistent

Accuracy/Dataset	Robotics	Flowers	Birds	Butterflies
Baseline (transductive)	5.51 ± 0.00	5.88 ± 0.00	23.33 ± 1.83	21.75 ± 0.00
Baseline (per-sample)	63.86 ± 0.41	72.82 ± 0.23	54.40 ± 0.34	53.90 ± 0.24
CoConut (unseen)	64.56 ± 0.48	74.88 ± 0.23	$\mathbf{54.87 \pm 0.35}$	54.03 ± 0.27
CoConut (structural)	–	–	–	–
CoConut (proportions)	$\mathbf{64.76 \pm 0.40}$	$\mathbf{75.15 \pm 0.22}$	54.53 ± 0.29	$\mathbf{54.42 \pm 0.34}$

Accuracy/Dataset	Cora	Citeseer
Baseline (transductive)	30.23 ± 0.00	20.18 ± 0.00
Baseline (per-sample)	69.05 ± 0.23	66.74 ± 0.18
CoConut (unseen)	–	–
CoConut (structural)	76.30 ± 0.47	$\mathbf{69.57 \pm 0.22}$
CoConut (proportions)	$\mathbf{77.58 \pm 0.33}$	68.31 ± 0.18