Discriminative Probabilistic Framework for Generalized Multi-Instance Learning

Anh T. Pham, Raviv Raich, Xiaoli Z. Fern, Weng K. Wong, and Xinze Guan

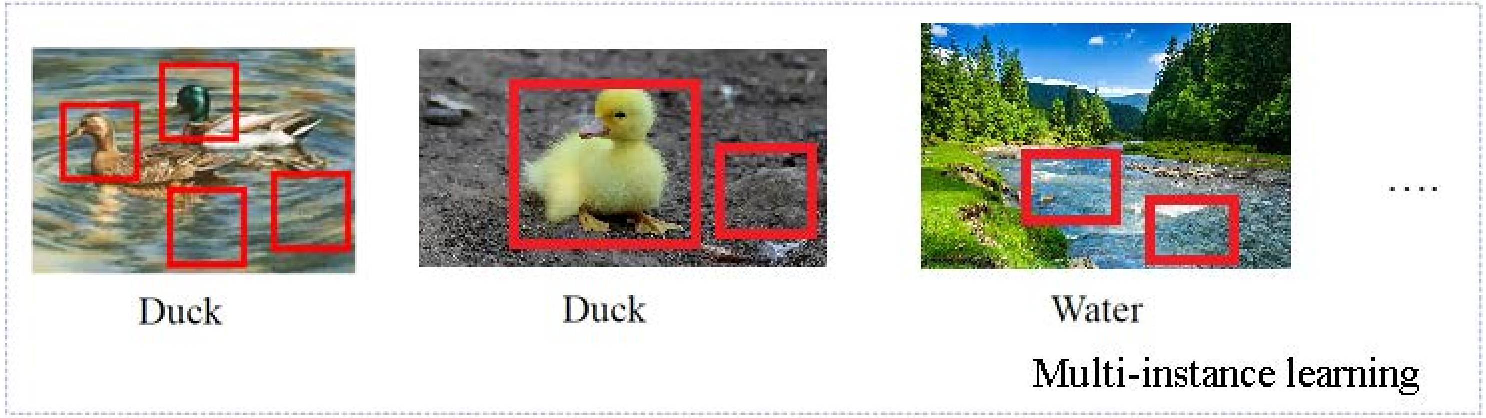
School of EECS, Oregon State University, Corvallis, OR 97331-5501, USA

{phaman,raich,xfern,wong}@eecs.oregonstate.edu and guanxinze@hotmail.com

Multiple-instance learning (MIL)

- Challenges with single-instance learning (SIL):
  - o Labeling is time-consuming (for big data)
  - o Difficult to obtain individual labels (privacy or prohibitive labeling costs)

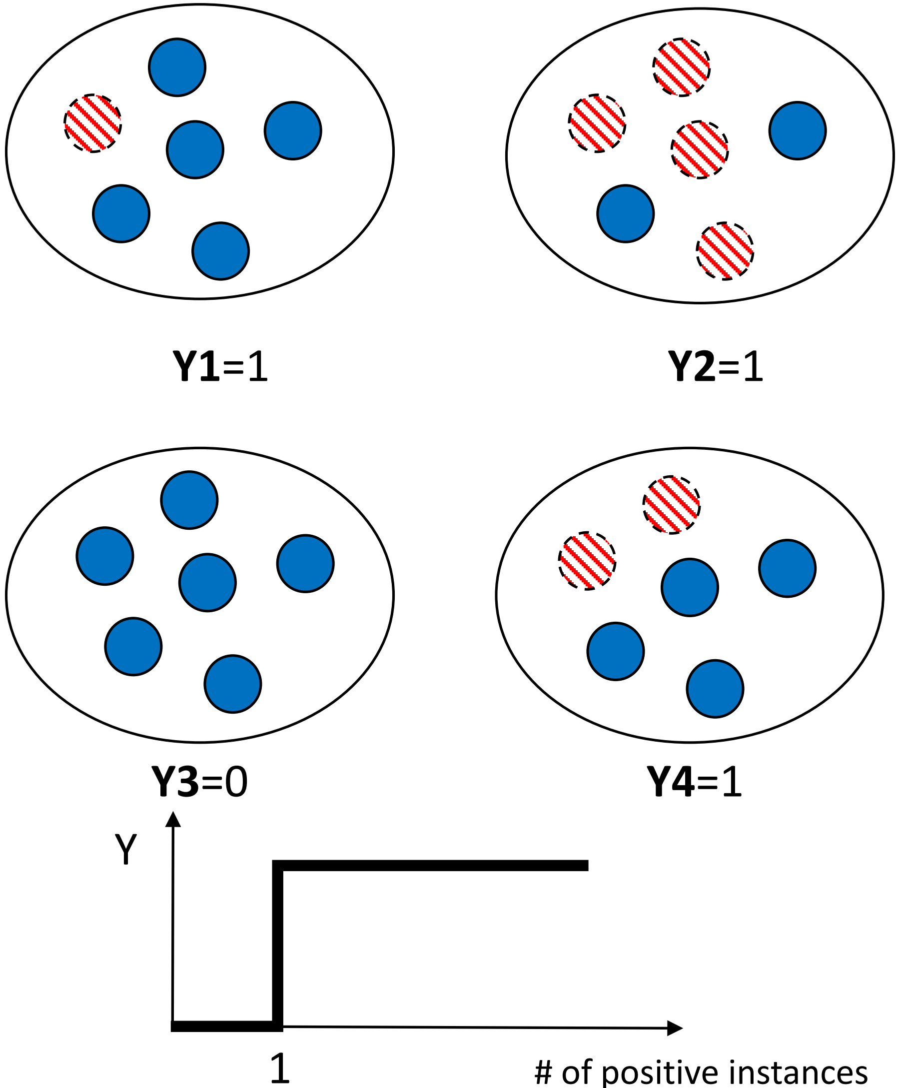
- Multi-instance learning: overcomes above labeling challenges



Example of generalized multi-instance learning

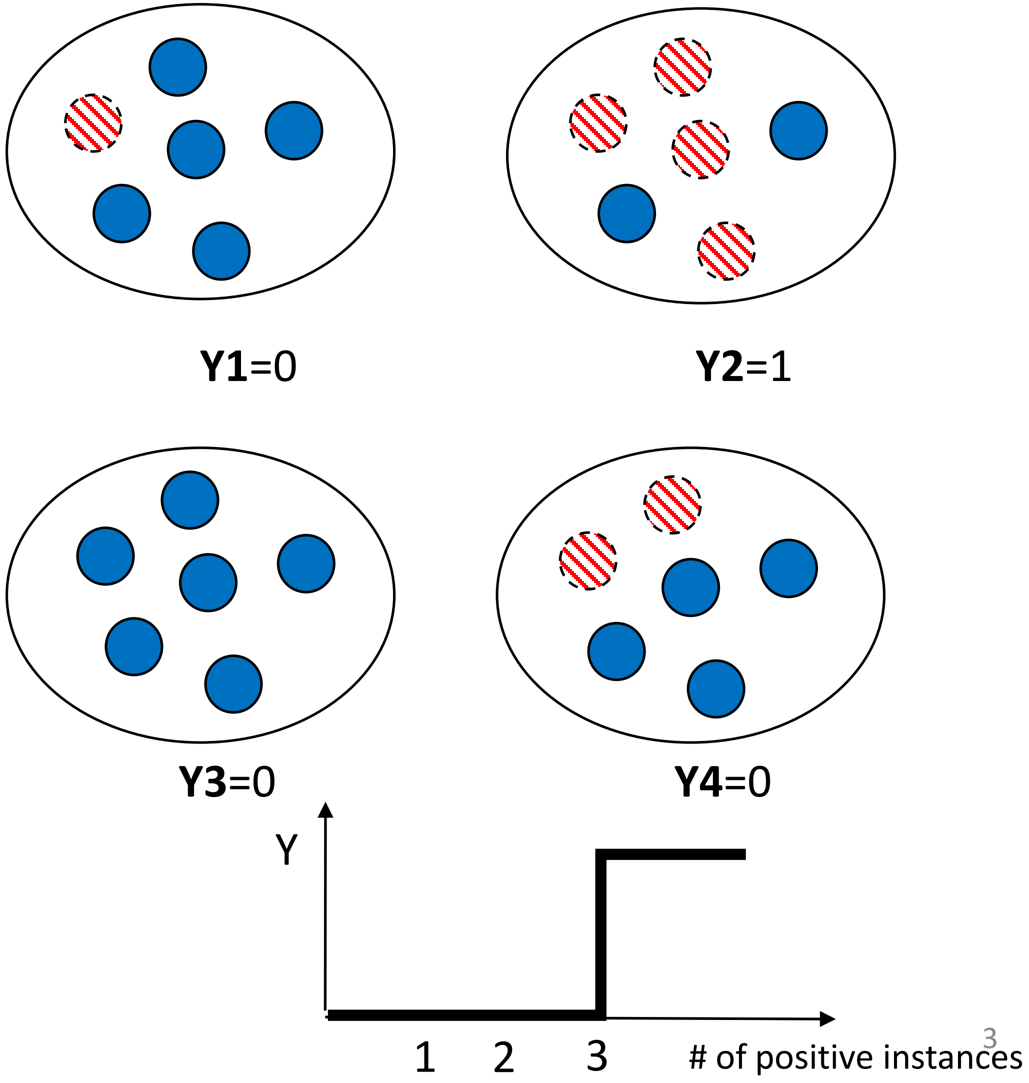
Presence-based assumption

Multi-instance learning

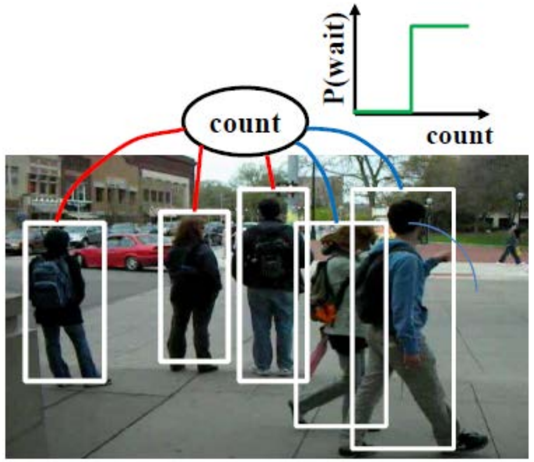


Count-based assumption

Generalized multi-instance learning

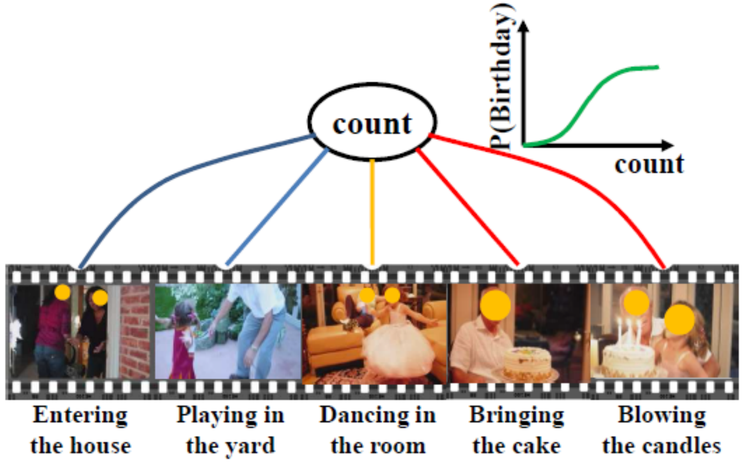


Applications of GMIL

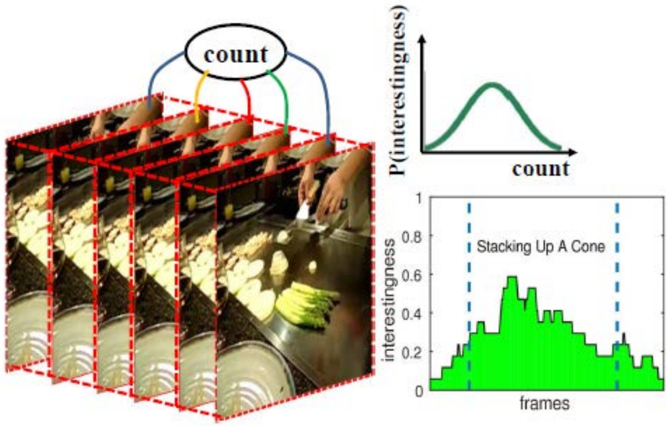


(a) What is the collective activity?

Figure 1. Example of images taken from Corel5k dataset. Images are labeled 'positive' w.r.t. the topic 'forest' if the image has more than ten 'tree's segments.



(b) What is the event of this video?



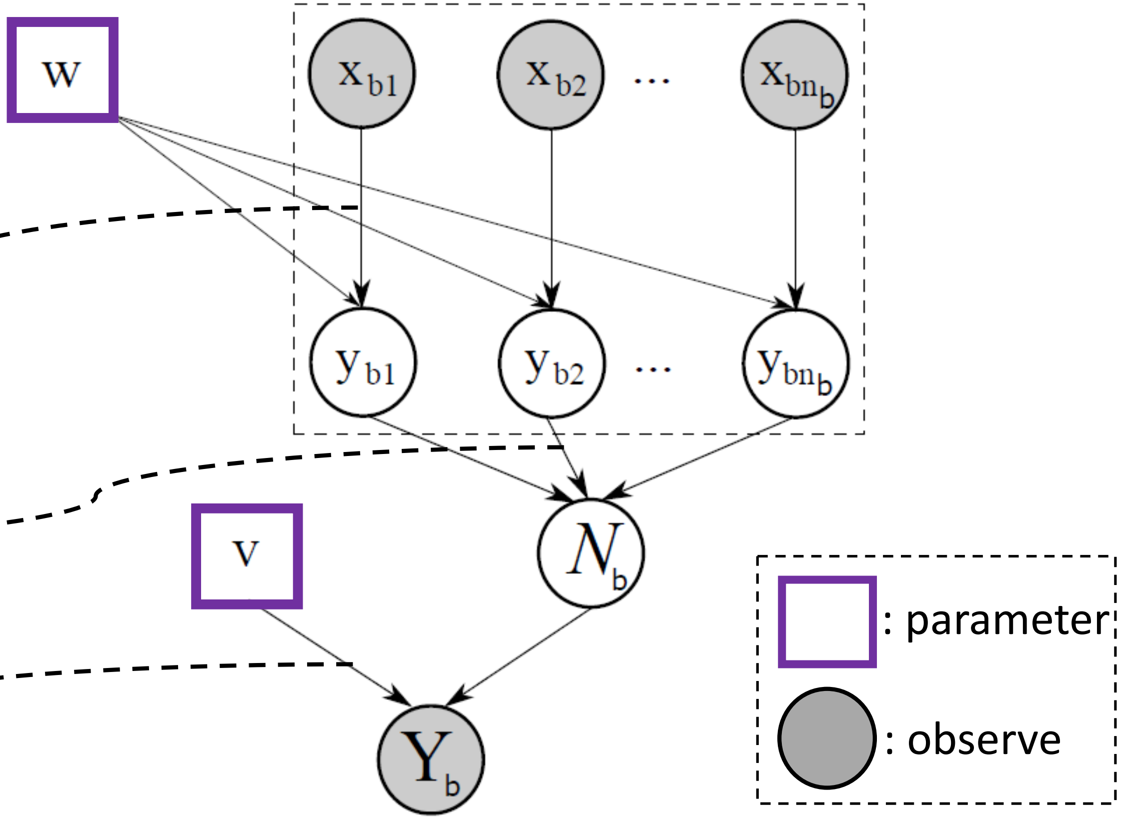
(c) Is this video segment interesting?

[1] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori. **Visual Recognition by Counting Instances: A Multi-Instance Cardinality Potential Kernel**. *(CVPR)*, 2015

Instance annotation in MIL

- **Training data:** B bags, represented by $\{(\mathbf{X}_b, \mathbf{Y}_b)\}_{b=1 \ldots n_b}$
  - $\mathbf{X}_b$ is a set of n_b instances for the b^{th} bag $\{\mathbf{x}_{b1}, \mathbf{x}_{b2}, \ldots, \mathbf{x}_{bn_b}\}$
    - Each instance $\mathbf{x}_{bi} \in R^d$ is associated with a label $y_{bi} \in \{0, 1\}$
  - $\mathbf{Y}_b \in \{0, 1\}$ is the bag label for the b^{th} bag
- **Goal:** Learn a classifier mapping $\mathbf{R}^d$ to {0 1}

Proposed model



Classifier parameter w

$$p(y_{bi} = 1|\mathbf{x}_{bi}, \mathbf{w}) = \frac{e^{\mathbf{w}^T \mathbf{x}_{bi}}}{1 + e^{\mathbf{w}^T \mathbf{x}_{bi}}}$$

$$p(N_b = n|\mathbf{y}_b) = \mathbb{I}[\sum_{i=1}^{n_b} y_{bi} = n]$$

Labeler parameter v

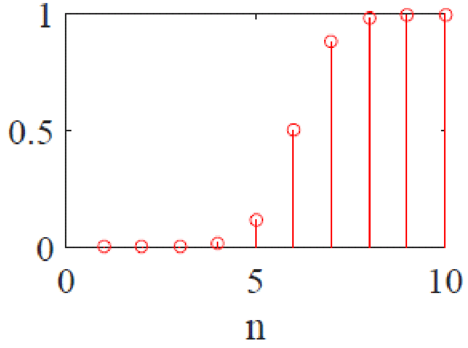
$$p(Y_b = 1|N_b = n, \mathbf{v}) = \phi(n; \mathbf{v})$$

Relation among:
-the bag label
-# of positive instances

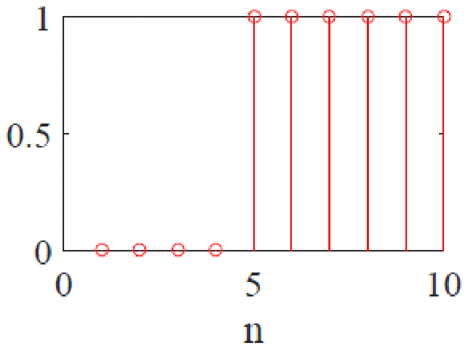
(Case 1) $\phi(n; \mathbf{v}) = \begin{cases} v_n & \text{if } n = 0, 1, \ldots, m, \\ v_m & \text{if } n > m. \end{cases}$

(Case 2) $\phi(n; \mathbf{v}) = \dfrac{e^{v_0 + n \times v_1}}{1 + e^{v_0 + n \times v_1}}$

(Case 3) $\phi(n; \mathbf{v}) = \mathbb{I}[n \geq v_0]$

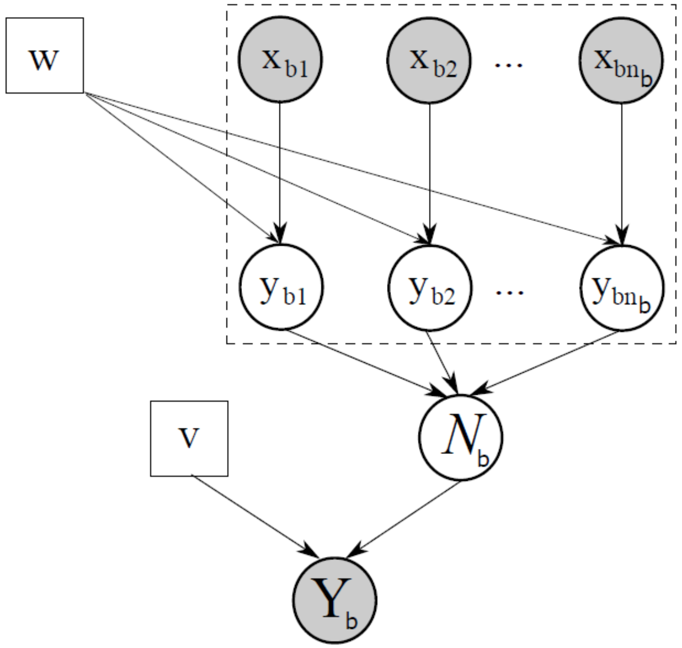


Case 2



Case 3

Goal

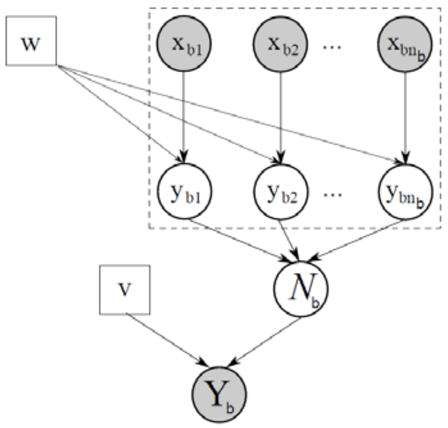


Our goal: To learn the instance level classifier parameter $\boldsymbol{w}$

Our challenge:

- Jointly learn the desired parameter $\boldsymbol{w}$ along with the nuisance parameter $\mathbf{v}$ of the bag labeler function $\phi(n; \mathrm{v})$.

EM for inference



Log-likelihood:

$$\mathbf{L}(\mathbf{Y}_D|\mathbf{X}_D,\mathbf{w},\mathbf{v}) = \sum_{b=1}^{B} \log[\sum_{N_b=0}^{n_b} \sum_{\mathbf{y}_b \subset M_b} p(Y_b, N_b, \mathbf{y}_b|\mathbf{X}_b,\mathbf{w},\mathbf{v})]$$

Complete log-likelihood:

$$\mathbf{L}(\mathbf{Y}_D,\mathbf{N}_D,\mathbf{y}_D|\mathbf{X}_D,\mathbf{w},\mathbf{v}) = \sum_{b=1}^{B}[\log p(Y_b|N_b,\mathbf{v}) + \log p(N_b|\mathbf{y}_b) + \log p(\mathbf{y}_b|\mathbf{x}_b,\mathbf{w})]$$

Surrogate function:

$$\mathcal{Q}(\mathbf{w},\mathbf{v};\mathbf{w}',\mathbf{v}') = \mathcal{Q}_1(\mathbf{w};\mathbf{w}',\mathbf{v}') + \mathcal{Q}_2(\mathbf{v};\mathbf{w}',\mathbf{v}')$$

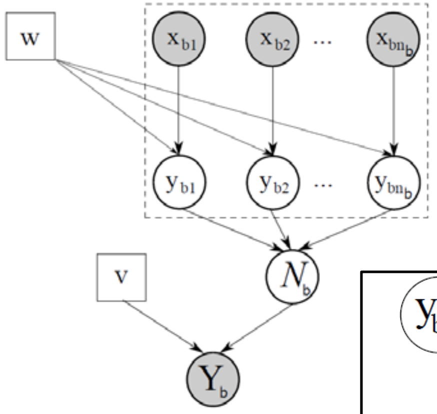
Depend on w only ← $\mathcal{Q}_1(\mathbf{w};\mathbf{w}',\mathbf{v}') = \sum_{b=1}^{B}\sum_{i=1}^{n_b}[p(y_{bi}=1|Y_b,\mathbf{X}_b,\mathbf{w}',\mathbf{v}')\mathbf{w}^T\mathbf{x}_{bi} - \log(1+e^{\mathbf{w}^T\mathbf{x}_{bi}})]$

E-step

Depend on v only ← $\mathcal{Q}_2(\mathbf{v};\mathbf{w}',\mathbf{v}') = \sum_{b=1}^{B}\sum_{n=0}^{n_b} p(N_b=n|Y_b,\mathbf{X}_b,\mathbf{w}',\mathbf{v}') \times (Y_b \log\phi(n;\mathbf{v}) + (1-Y_b)\log(1-\phi(n;\mathbf{v})))$

separate M-step for w, v

E-step



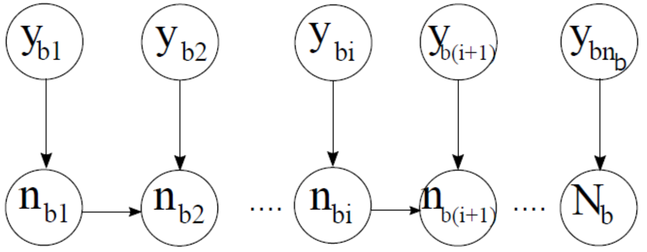
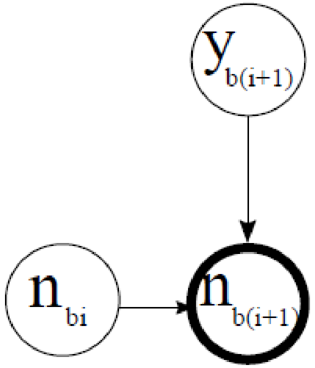


Figure 4. Graphical model to compute instance membership probability $p(y_{bi} = 1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ and probability of the number of positive instances $p(N_b = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$.

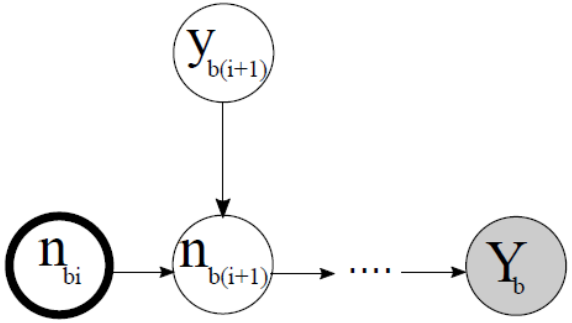
$$n_{bi} = \sum_{j=1}^{i} y_{bj} \qquad n_{b(i+1)} = n_{bi} + y_{b(i+1)}$$



Forward message

$$\alpha_i(k) \triangleq p(n_{bi} = k|\mathbf{X}_b, \mathbf{w}')$$

$$\alpha_{i+1}(k) = p(y_{b(i+1)} = 1|\mathbf{x}_{bi}, \mathbf{w}')\alpha_i(k-1) + p(y_{b(i+1)} = 0|\mathbf{x}_{bi}, \mathbf{w}')\alpha_i(k)$$



Backward message

$$\beta_i(k) \triangleq p(Y_b|n_{bi} = k, \mathbf{X}_b, \mathbf{w}')$$

$$\beta_i(k) = p(y_{b(i+1)} = 1|\mathbf{x}_{bi}, \mathbf{w}')\beta_{i+1}(k+1) + p(y_{b(i+1)} = 0|\mathbf{x}_{bi}, \mathbf{w}')\beta_{i+1}(k)$$

E-step 1

$$p(y_{bi} = 1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}') = \frac{\pi_1}{\pi_1 + \pi_0}$$

$$\pi_1 = \sum_{k=0}^{n_b} \alpha_{i-1}(k)\beta_i(k+1)p(y_{bi} = 1|\mathbf{x}_{bi}, \mathbf{w}') \,;$$

$$\pi_0 = \sum_{k=0}^{n_b} \alpha_{i-1}(k)\beta_i(k)p(y_{bi} = 0|\mathbf{x}_{bi}, \mathbf{w}')$$

E-step 2

$$p(n_{bi} = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}') = \frac{\alpha_i(k)\beta_i(k)}{\sum_{k=0}^{n_b} \alpha_i(k)\beta_i(k)}$$

$p(N_b = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ is obtained for $i = n_b$

M-step: Update $\mathbf{w}$

$$\mathcal{Q}(\mathbf{w},\mathbf{v};\mathbf{w}',\mathbf{v}') = \mathcal{Q}_1(\mathbf{w};\mathbf{w}',\mathbf{v}') + \mathcal{Q}_2(\mathbf{v};\mathbf{w}',\mathbf{v}')$$

$$\mathcal{Q}_1(\mathbf{w};\mathbf{w}',\mathbf{v}') = \sum_{b=1}^{B}\sum_{i=1}^{n_b}[p(y_{bi}=1|Y_b,\mathbf{X}_b,\mathbf{w}',\mathbf{v}')\mathbf{w}^T\mathbf{x}_{bi} - \log(1+e^{\mathbf{w}^T\mathbf{x}_{bi}})]$$

$$\mathcal{Q}_2(\mathbf{v};\mathbf{w}',\mathbf{v}') = \sum_{b=1}^{B}\sum_{n=0}^{n_b} p(N_b=n|Y_b,\mathbf{X}_b,\mathbf{w}',\mathbf{v}')\times (Y_b\log\phi(n;\mathbf{v}) + (1-Y_b)\log(1-\phi(n;\mathbf{v})))$$

→ Can update $\mathbf{w}$ and $\mathbf{v}$ separately

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \eta\times\mathbf{H}_w^{-1}\mathbf{d}_w\Big|_{\mathbf{w}=\mathbf{w}^{(k)}}$$

$$\mathbf{d}_w = \sum_{b=1}^{B}\sum_{i=1}^{n_b}[p(y_{bi}=1|Y_b,\mathbf{X}_b,\theta') - p(y_{bi}=1|\mathbf{x}_{bi},\mathbf{w})]\mathbf{x}_{bi}$$

$$\mathbf{H}_w = -\sum_{b=1}^{B}\sum_{i=1}^{n_b} p(y_{bi}=1|\mathbf{x}_{bi},\mathbf{w})p(y_{bi}=0|\mathbf{x}_{bi},\mathbf{w})\mathbf{x}_{bi}\mathbf{x}_{bi}^T$$

Newton method to update $\mathbf{w}$

M-step: Update $\mathbf{v}$

(Case 1) $\phi(n;\mathbf{v}) = \begin{cases} v_n & \text{if } n = 0, 1, \ldots, m, \\ v_m & \text{if } n > m. \end{cases}$

(Case 2) $\phi(n;\mathbf{v}) = \dfrac{e^{v_0 + n \times v_1}}{1 + e^{v_0 + n \times v_1}}$

(Case 3) $\phi(n;\mathbf{v}) = \mathbb{I}[n \geq v_0]$

Case 1

$$v_n = \frac{\delta_n}{\delta_n + \tau_n},$$

$$\delta_n = \sum_{b=1}^{B} Y_b p(N_b = n | Y_b = 1, \mathbf{X}_b, \theta')$$

$$\tau_n = \sum_{b=1}^{B} (1 - Y_b) p(N_b = n | Y_b = 0, \mathbf{X}_b, \theta')$$

Closed form

Case 2

$$\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} - \eta \times \mathbf{H}_v^{-1} \mathbf{d}_v \Big|_{\mathbf{v}=\mathbf{v}^{(k)}}$$

Newton method

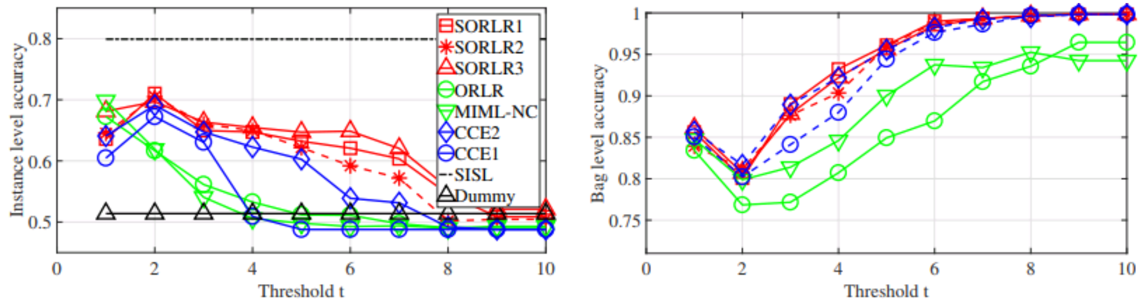
Case 3

$$\min_{v_0} \sum_{b=1}^{B} [Y_b p(N_b < v_0 | Y_b = 1, \mathbf{X}_b, \theta') + (1 - Y_b) p(N_b \geq v_0 | Y_b = 0, \mathbf{X}_b, \theta')]$$

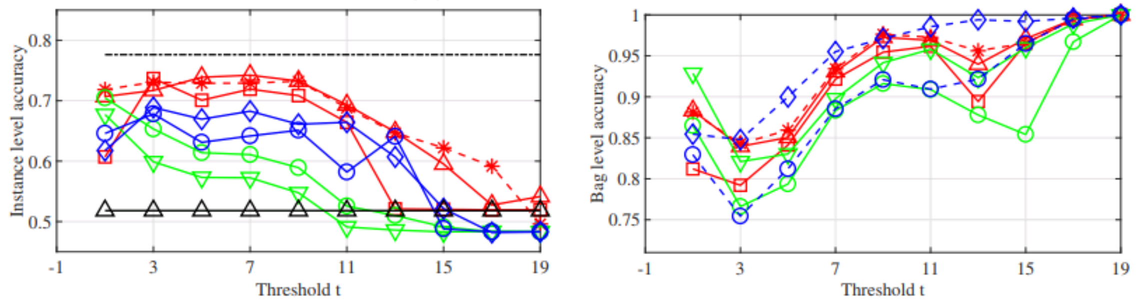
v_0 is discrete

Real dataset experiments

- **Datasets.**
  - HJA bird song recording, MSCV2 image annotation
  - Select some classes as positive and treat remaining as negative
  - Bag label is generated with threshold t
- **Settings.**
  - **SISL classifier**: all instance labels are known. **Dummy classifier**: classifies instances purely based on label proportions
  - **CCE**: Constructive Clustering Ensemble (SVM-based approach)
    - CCE1: does not take into positive instances count
    - CCE2: does take into positive instances count
  - **SORLR1-3:** Implement for 3 cases
  - **ORLR:** v_0 is set to 0
  - **MIMLNC:** negative class is considered novel
- **Results and Analysis.**
  - Consider both instance level prediction and bag level predictions
  - SORLR outperforms other baseline, especially SORLR3 since it models the data quite well.
  - When t is large, most bags are negative. Thus, bag level prediction is high, but instance prediction is low



(a) MSCV2 dataset



(b) HJA dataset

Conclusions

- A discriminative probabilistic model for generalized multi-instance learning is presented
- An exact and efficient inference framework is proposed
- Experiments show promising results compared to state-of-the-art alternatives