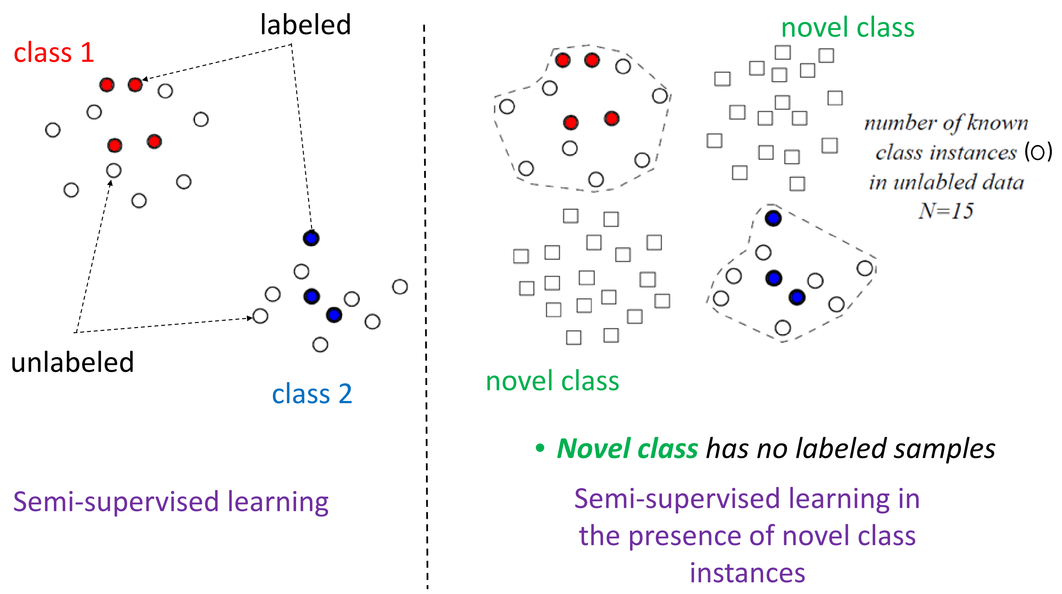
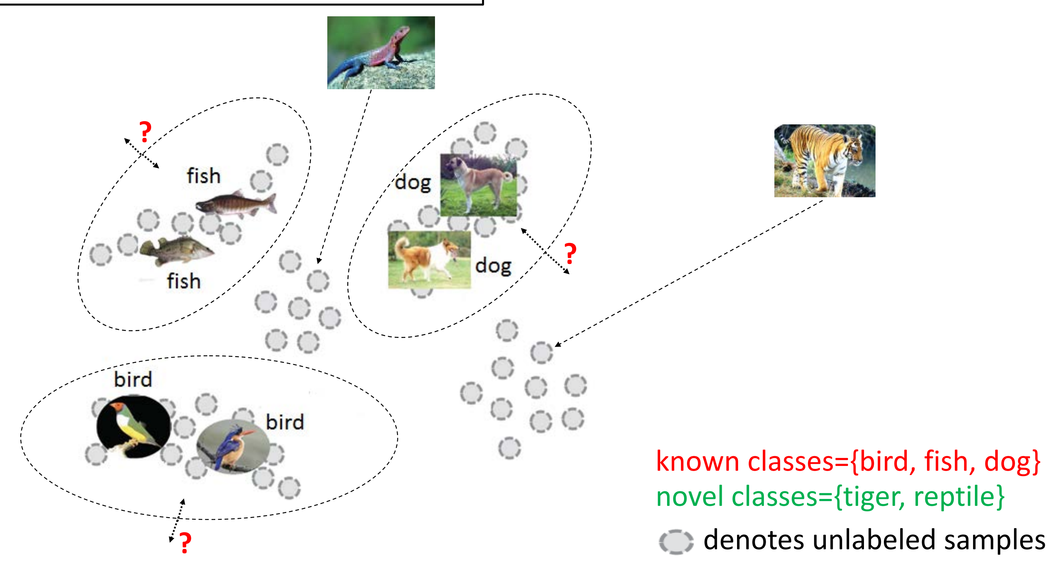
Semi-supervised learning in the presence of novel class instances

Anh T. Pham, Raviv Raich, and Xiaoli Z. Fern
School of EECS, Oregon State University, Corvallis, OR 97331-5501, USA
{phaman,raich,xfern}@eecs.oregonstate.edu

1 Semi-supervised learning with novel class instances



2 Examples in computer vision



Goal: Obtain margins for {bird, fish, dog} and remaining regions for {novel class}

3 Problem formulation

Training data: Labeled data $\{\mathbf{x}_i, y_i\}_{i=1}^{L}$ and unlabeled data $\{\mathbf{x}_{L+i}\}_{i=1}^{U}$

1. Labeled data: $\mathbf{x}_i \in \underline{X}=R^d$, $y_i \in \{1, \ldots, C\}$ which are known classes.
2. Unlabeled data: $\mathbf{x}_{L+i} \in R^d$, y_{L+i} is unknown and $y_{L+i} \in \{0, 1, \ldots, C\}$ which are both known classes and novel class (denoted as 0).
3. Number of novel class instances: N.

Goal: Find a classifier that maps an instance in $\underline{X}$ into a label in $\underline{Y}=\{0,1, \ldots, C\}$

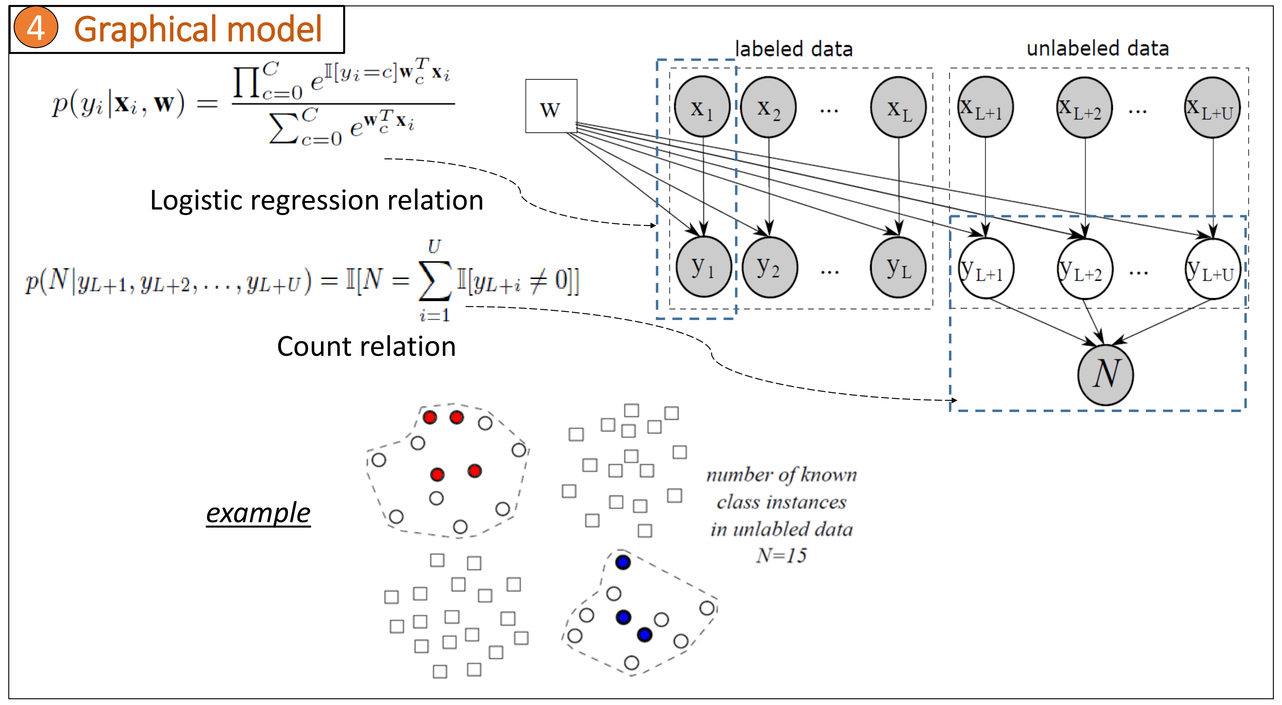
4 Graphical model

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{\prod_{c=0}^{C} e^{\mathbb{I}[y_i=c]\mathbf{w}_c^T\mathbf{x}_i}}{\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_i}}$$

Logistic regression relation

$$p(N|y_{L+1}, y_{L+2}, \ldots, y_{L+U}) = \mathbb{I}[N = \sum_{i=1}^{U} \mathbb{I}[y_{L+i} \neq 0]]$$

Count relation



example

5 Inference

The log-likelihood $\mathbf{L}(\mathbf{w}) = \sum_{i=1}^{L} \log p(y_i|\mathbf{x}_i, \mathbf{w}) + \log p(N|\mathbf{X}_U, \mathbf{w})$

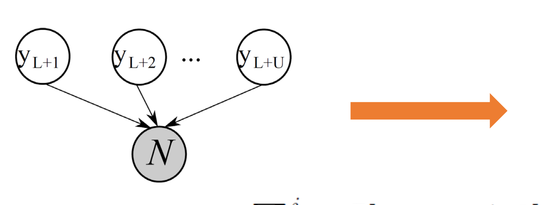
The surrogate function

$$\begin{aligned} g(\mathbf{w}, \mathbf{w}') &= E_{\mathbf{y}_U|\mathbb{D},\mathbf{w}'} \log p(\mathbf{y}_L, \mathbf{y}_U, N|\mathbf{X}_L, \mathbf{X}_U, \mathbf{w}) \\ &= E_{\mathbf{y}_U|\mathbb{D},\mathbf{w}'} [\log p(\mathbf{y}_L|\mathbf{X}_L, \mathbf{w}) + \log p(\mathbf{y}_U|\mathbf{X}_U, \mathbf{w})] \\ &\quad + E_{\mathbf{y}_U|\mathbb{D},\mathbf{w}'} \log p(N|\mathbf{y}_U, \mathbf{X}_U, \mathbf{w}). \end{aligned}$$

$$g(\mathbf{w}, \mathbf{w}') = \sum_{i=1}^{L}[\sum_{c=0}^{C} \mathbb{I}[y_i = c]\mathbf{w}_c^T\mathbf{x}_i - \log(\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_i})] +$$

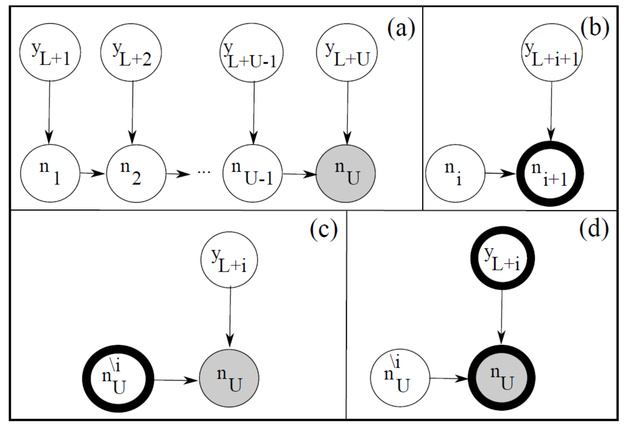
$$\sum_{i=1}^{U}\sum_{c=0}^{C} p(y_{L+i} = c|N, \mathbf{X}_U, \mathbf{w}')[\mathbf{w}_c^T\mathbf{x}_{L+i} - \log(\sum_{l=0}^{C} e^{\mathbf{w}_l^T\mathbf{x}_{L+i}})]$$

6 E-step: Compute $p(y_{L+i} = c|N, \mathbf{X}_U, \mathbf{w})$



$n_i = \sum_{k=1}^{i} \mathbb{I}[y_{L+k} \neq 0]$

Number of known class instances from the 1 to ith unlabeled instances



1. Compute $p(n_U|\mathbf{X}_U, \mathbf{w})$
2. Compute $p(n_U^{\setminus i}|\mathbf{X}_U, \mathbf{w})$
3. Compute $p(y_{L+i} = c, n_U = N|\mathbf{X}_U, \mathbf{w})$ then $p(y_{L+i} = c|N, \mathbf{X}_U, \mathbf{w})$

7 2D Toy dataset results

Datasets:

- Instances are uniformly generated from 6 areas representing 5 classes (Fig. 4(a))
- Novel class is denoted as 0
- 100 labeled samples and 1000 unlabeled samples (Fig. 4(b))
- There are **667** known class instances in the unlabeled data
- Perform the proposed approach with N decreases from **950 → 500**

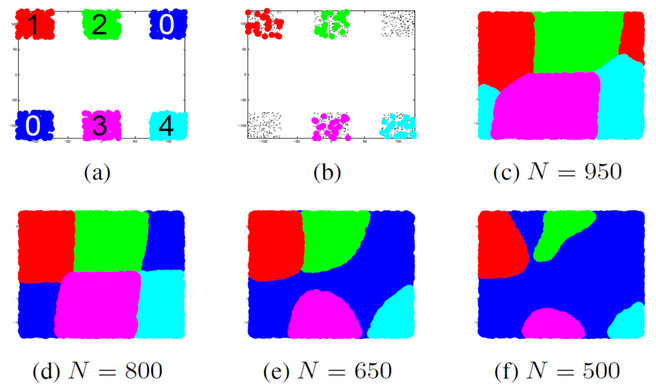


Fig. 4: Results on a toy dataset. (a) Class distribution, novel class instances are blue. (b) Labeled instances (red, pink, green, and cyan dots) and unlabeled instances (small black dots). (c-f) Learned boundary when $N = 950$, $N = 800$, $N = 650$, and $N = 500$, respectively.

8 Real dataset results

Baseline methods:

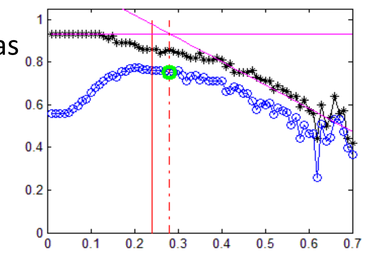
1. LRSSS-opt: proposed method
2. LRSSS-tune: proposed method with tuned N
3. LACU [1]
4. LR-L: trained with labeled data only
5. LRSSS-true: trained with labeled data only

Datasets:

HJA bird [2], MSCV2, 50Salad, MNIST: Using the default class order: class1, 2, 3 are known, class4 is novel

Tuning N:

- Select labeled data as validation set
- Accuracy vs N curve
- Detect the knee



Results:

Dataset	HJA bird	MSCV2	50Salad	MNIST
LRSSS-opt	**74.4±1.7**	**77.8±2.2**	**72.8±2.2**	**82.0±3.1**
LRSSS-tune	73.2±2.5	69.0±1.5	69.0±2.7	69.2±3.3
LACU	51.6±8.8	65.8±7.1	65.7±7.5	79.4±3.3
LR-L	54.1±2.1	73.2±1.7	42.1±1.9	67.7±2.4
LRSSS-true	74.3±1.3	78.0±2.1	70.1±2.5	78.5±3.3

Table 1: Accuracy results of the proposed approach and LACU. The proposed method and indistinguishable values using 95% confidence two-tailed paired *t*-tests with the highest values are bolded.

9 Conclusion

- A framework for semi-supervised learning in the presence of novel class instances
- A probabilistic approach to control the cardinality the of known class instances in the unlabeled data
- Future work:
  - Estimate the number of novel class instances
  - Improve the computational efficiency for the E-step

References

[1] Q. Da, Y. Yu, and Z.-H. Zhou, "Learning with augmented class by exploiting unlabeled data," in AAAI Conference on Artificial Intelligence, 2014, pp. 1760–1766.

Acknowledgements: This work is partially supported by the National Science Foundation grants CCF-1254218, DBI-1356792, and IIS-1055113.

ICASSP 2016