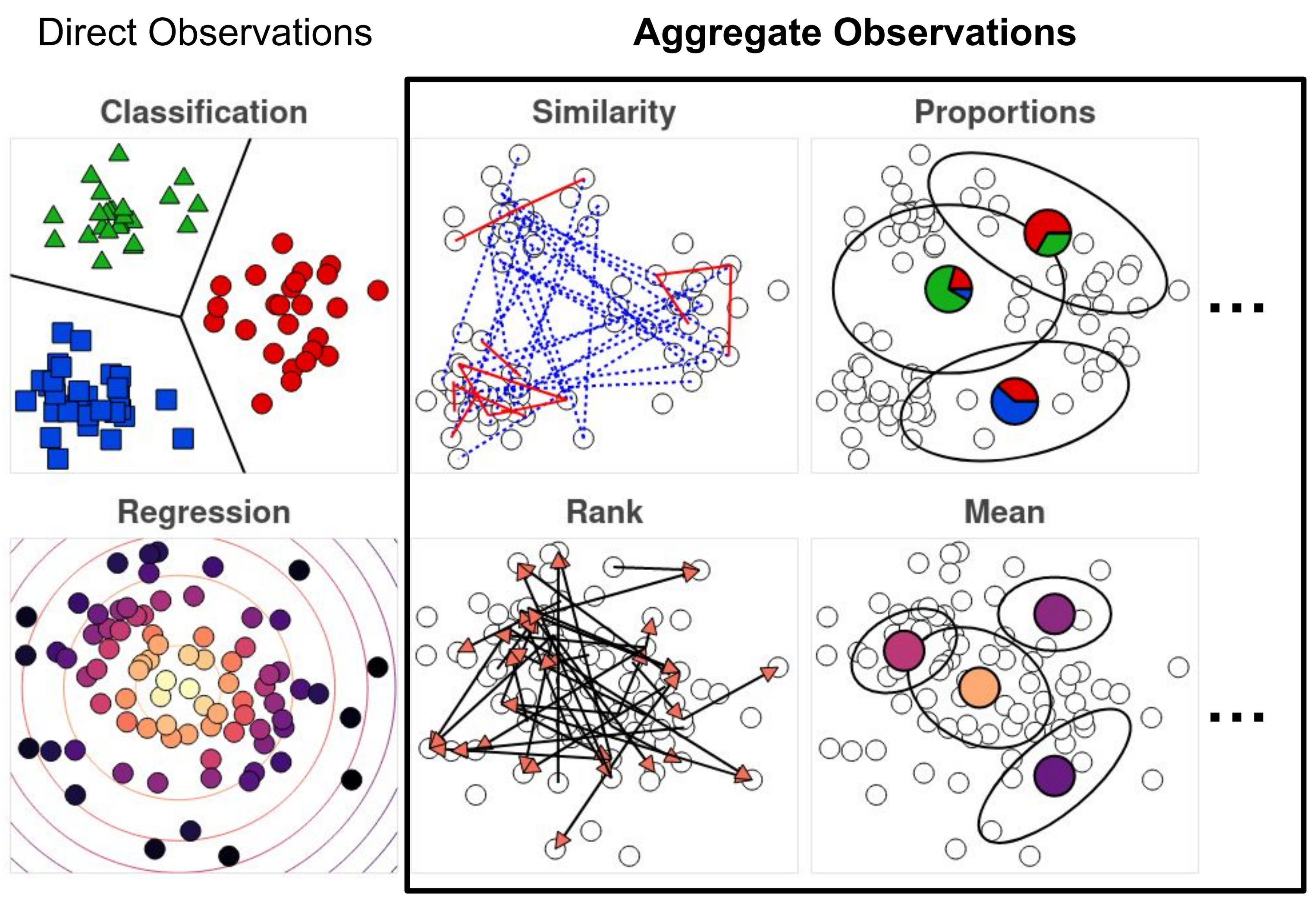
Learning from Aggregate Observations

Yivan Zhang[1, 2], Nontawat Charoenphakdee[1, 2], Zhenguo Wu[1], Masashi Sugiyama[2, 1]
[1]The University of Tokyo, [2]RIKEN

Introduction



Motivation scarcity of individual labels Schuessler (1999), Zhou (2004, 2018)

- **Expensive**: video annotation; semantic segmentation
- **Privacy sensitive**: census, medical or public health data analysis
- **Intrinsically unavailable**: drug activity prediction; remote sensing

Data supervision given to **sets of instances**

Task to predict labels of **individual instances**

Related work

- **Multiple Instance Learning (MIL)** Zhou (2004) and **Learning from Label Proportions (LLP)** Kück+ (2005): only for binary classification
- **Classification via pairwise similarity** Hsu+ (2019): our special case

Our contribution

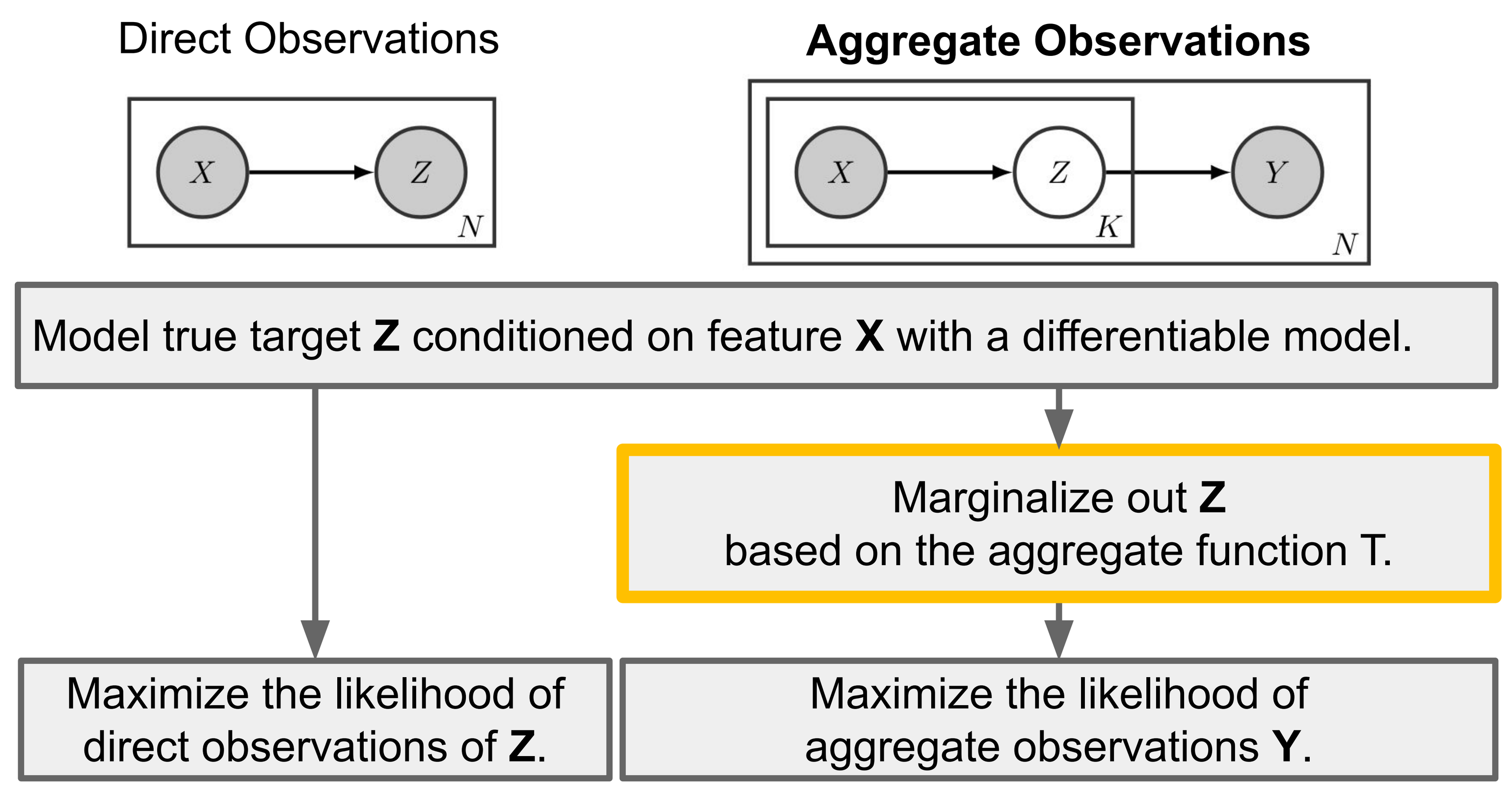
- A general probabilistic framework for aggregate observations for classification and regression problems
- A simple method applicable to any differentiable models such as deep neural networks and gradient boosting machines
- Theoretical justification based on the concept of consistency up to an equivalence relation

Examples

Learning from …	Aggregate Observation $Y = T(Z_{1:K})$
similarity/dissimilarity ($K = 2$)	if Z_1 and Z_2 are the same or not
triplet comparison ($K = 3$)	if $d(Z_1, Z_2)$ is smaller than $d(Z_1, Z_3)$, where $d(\cdot,\cdot)$ is a similarity measure between classes
multiple instance ($K \geq 2$)	if $Z_{1:K}$ contains positive instances ($C = 2$)
mean/sum ($K \geq 2$)	the arithmetic mean or the sum of $Z_{1:K}$
difference/rank ($K = 2$)	the difference $Z_1 - Z_2$, or the relative order $Z_1 > Z_2$
min/max ($K \geq 2$)	the smallest/largest value in $Z_{1:K}$
uncoupled data ($K \geq 2$)	randomly permuted $Z_{1:K}$

input feature $X \in \mathcal{X}$ **true target** $Z \in \mathcal{Z}$ **aggregate observation** $Y \in \mathcal{Y}$
aggregate function $T : \mathcal{Z}^K \mapsto \mathcal{Y}$, i.e., $Y = T(Z_{1:K})$

Proposed Method



Aggregate observation assumption

True targets contain all information to predict the aggregate observation.

$$p(Y|X_{1:K}, Z_{1:K}) = p(Y|Z_{1:K})$$

Independent observations assumption

True targets are mutually independent in sets.

This assumption may be violated in real-world applications.

$$p(Z_{1:K}|X_{1:K}) = \prod_{i=1}^{K} p(Z_i|X_i)$$

Joint probability factorization

$$p(X_{1:K}, Z_{1:K}, Y) = p(Y|Z_{1:K}) \prod_{i=1}^{K} p(Z_i|X_i)p(X_i)$$

Marginalization over Z

- **Classification**: summation → always analytically calculable
- **Regression**: depending on the aggregate function and distribution

$$p(Y|X_{1:K}) = \int_{\mathcal{Z}^K} \delta_{T(z_{1:K})}(Y) \prod_{i=1}^{K} p(z_i|X_i)\, \mathrm{d}z_{1:K} = \mathop{\mathbb{E}}_{\substack{Z_i \sim p(Z_i|X_i) \\ i=1,\dots,K}} \left[\delta_{T(Z_{1:K})}(Y)\right]$$

Log-likelihood of Y

$$\ell_N(W) = \frac{1}{N} \sum_{i=1}^{N} \log p(y^{(i)}|x_{1:K}^{(i)}; W)$$

Realizations

Pairwise Similarity Hsu+ (2019)

$$Y = T_{\mathrm{sim}}(Z_1, Z_2) = [Z_1 = Z_2]$$

$$p(Y = 1) = \sum_{i=1}^{C} p(Z_1 = i)p(Z_2 = i)$$

Triplet Comparison

$$Y = T_{\mathrm{tri}}(Z_1, Z_2, Z_3) = [d(Z_1, Z_2) < d(Z_1, Z_3)]$$

$$p(Y = 1) = \sum_{\substack{d(i,j)<d(i,k) \\ i,j,k\in\{1,\dots,C\}}} p(Z_1 = i)p(Z_2 = j)p(Z_3 = k)$$

Mean Observation

$$Y = T_{\mathrm{mean}}(Z_{1:K}) = \frac{1}{K} \sum_{i=1}^{K} Z_i$$

$$Y \sim \mathcal{N}\left(\frac{1}{K} \sum_{i=1}^{K} \mu_i, \frac{1}{K^2} \sum_{i=1}^{K} \sigma_i^2\right)$$

Rank Observation

$$Y = T_{\mathrm{rank}}(Z_1, Z_2) = [Z_1 > Z_2]$$

$$p(Z_1 > Z_2) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mu_1 - \mu_2}{\sqrt{2(\sigma_1^2 + \sigma_2^2)}}\right)\right]$$

Other distributions:

Poisson - count data **Cauchy** - robust regression **Gumbel** - extreme value

Consistency up to an Equivalence Relation

Aggregate observations may not contain all information about individuals.

How much individual information is learned from aggregate observations?

Definition 3 (Equivalence). An *equivalence relation* $\sim$ on $\mathcal{W}$ induced by the likelihood is defined according to $W \sim W' \iff \ell(W) = \ell(W')$. The equivalence class of W is denoted by $[W]$.

Partially identifiable model: equivalent parameters → equal likelihood

Definition 4 (Consistency up to $\sim$). An estimator $\widehat{W}_N$ is said to be *consistent up to an equivalence relation* $\sim$, if $d(\widehat{W}_N, [W_0]) \xrightarrow{\mathrm{p}} 0$ as $N \to \infty$, where $d(W, [W_0]) = \inf_{W_0' \in [W_0]} d(W, W_0')$.

An estimator that converges to a value that is equivalent to the true one

Examples:

- Classification via pairwise similarity/triplet comparison is *at most* consistent up to a permutation
- Regression via mean observation is consistent
- Regression via rank observation is consistent up to an additive constant

If the estimator is only consistent up to an equivalence relation:

- Obtain only **partial information** about individual labels
- Incorporate easier-to-obtain **side information** about data
- Combine other sources of **strong/weak supervision**

Experiments

- Classification via **pairwise similarity** and **triplet comparison** on MNIST/Fashion-MNIST/Kuzushiji-MNIST datasets (CNN models)
- Evaluation: optimal permutation + accuracy Hsu+ (2019)

Dataset	Unsupervised	Pairwise Similarity			Triplet Comparison			Supervised
		Siamese	Contrastive	Ours/Hsu+	Tuplet	Triplet	Ours	
MNIST	52.30	**85.82**	98.45	**98.84**	18.42	22.77	**94.94**	99.04
	(1.15)	**(24.86)**	(0.11)	**(0.10)**	(1.08)	(9.38)	**(3.68)**	(0.08)
Fashion-MNIST	50.94	62.86	88.49	**90.59**	21.98	27.27	**81.49**	91.97
	(3.28)	(17.97)	(0.28)	**(0.26)**	(0.72)	(12.82)	**(0.94)**	(0.24)
Kuzushiji-MNIST	40.22	61.30	89.65	**93.45**	16.00	20.39	**81.94**	94.47
	(0.01)	(17.41)	(0.19)	**(0.32)**	(0.27)	(2.03)	**(4.59)**	(0.21)

- Regression via **mean observation** and **rank observation** on UCI datasets (linear regression & gradient boosting machines)
- Evaluation: optimal constant shift + MSE → error variance

Dataset	Mean Observation				Rank Observation				Supervised	
	Baseline		Ours		RankNet, Gumbel		Ours, Gaussian			
	LR	GBM	LR	GBM	LR	GBM	LR	GBM	LR	GBM
abalone	7.91	7.89	5.27	**4.80**	5.81	10.66	**5.30**	**5.04**	5.00	4.74
	(0.4)	(0.5)	(0.4)	**(0.3)**	(0.4)	(0.7)	**(0.3)**	**(0.5)**	(0.3)	(0.4)
airfoil	38.57	28.65	23.59	**4.63**	37.15	47.46	27.95	**6.18**	22.59	3.84
	(2.0)	(2.5)	(1.8)	**(0.9)**	(1.8)	(3.7)	(1.1)	**(1.0)**	(1.9)	(0.5)
auto-mpg	41.59	36.31	14.61	**9.53**	27.26	65.39	17.34	**9.97**	11.73	7.91
	(5.7)	(1.9)	(3.2)	**(2.4)**	(4.0)	(7.4)	(2.0)	**(2.0)**	(2.3)	(1.6)
concrete	198.51	172.35	115.06	**31.84**	244.06	268.86	233.93	**38.11**	111.92	24.80
	(12.8)	(15.2)	(10.1)	**(3.0)**	(17.1)	(26.5)	(20.0)	**(5.4)**	(6.4)	(5.7)
housing	67.40	52.23	27.54	**14.85**	52.51	93.07	44.40	**23.49**	29.66	13.12
	(20.8)	(6.0)	(6.8)	**(3.0)**	(10.8)	(8.1)	(13.4)	**(6.9)**	(6.1)	(3.7)
power-plant	172.64	170.10	20.73	**12.82**	163.64	294.07	44.82	**26.06**	21.17	11.84
	(7.1)	(3.4)	(0.8)	**(0.6)**	(4.8)	(4.9)	(6.1)	**(2.5)**	(1.0)	(0.9)

References

Alexander A Schuessler. "Ecological inference." Proceedings of the National Academy of Sciences, 96.19: 10578–10581, 1999.

Zhi-Hua Zhou. "Multi-instance learning: A survey." Nanjing University, Tech. Rep, 2004.

Hendrik Kück and Nando de Freitas. "Learning about individuals from group statistics." Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, 2005.

Zhi-Hua Zhou. "A brief introduction to weakly supervised learning." National Science Review, 5.1: 44–53, 2018.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. "Multi-class classification without multi-class labels." In International Conference on Learning Representations, 2019.