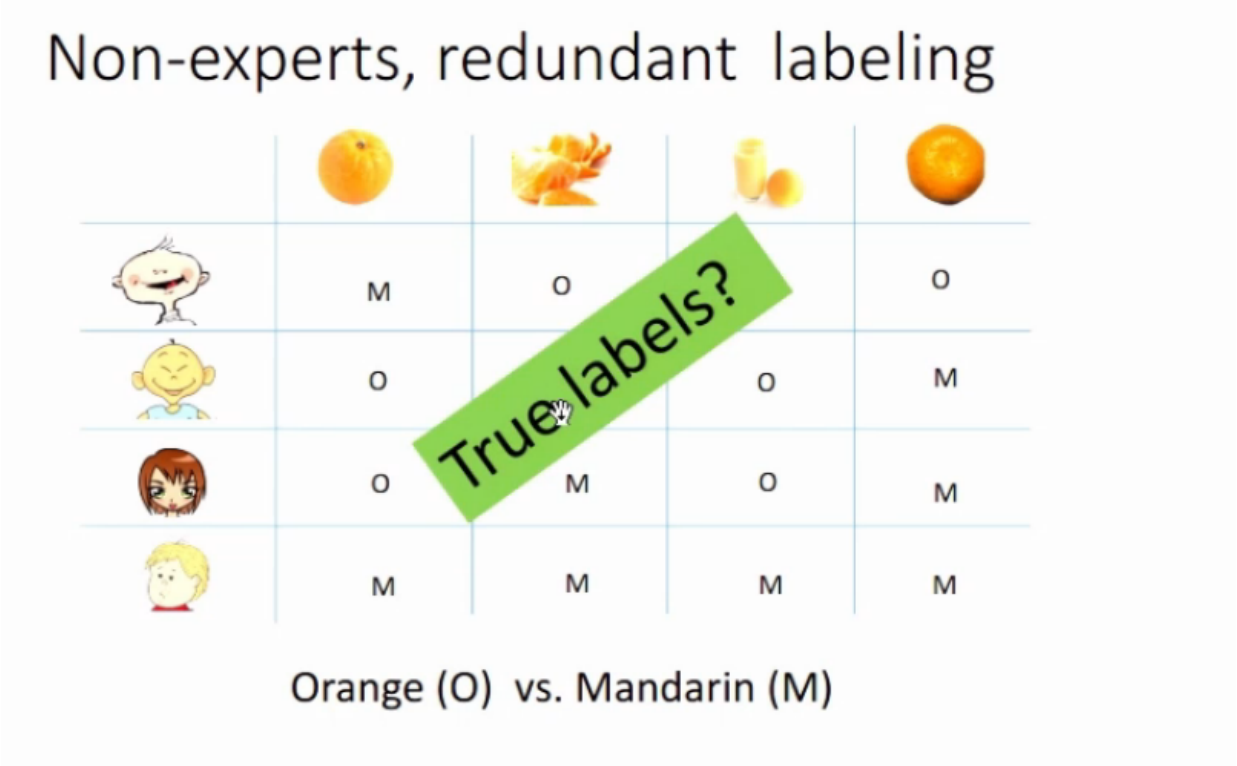
Joint Generative-Discriminative Aggregation Model for Multi-Option Crowd Labels

Kamran Ghasedi Dizaji[†], Yanhua Yang[‡], Heng Huang[†]

[†]**Electrical and Computer Engineering Department, University of Pittsburgh, USA**
[‡]**School of Electronic Engineering, Xidian University, China**

WSDM 2018

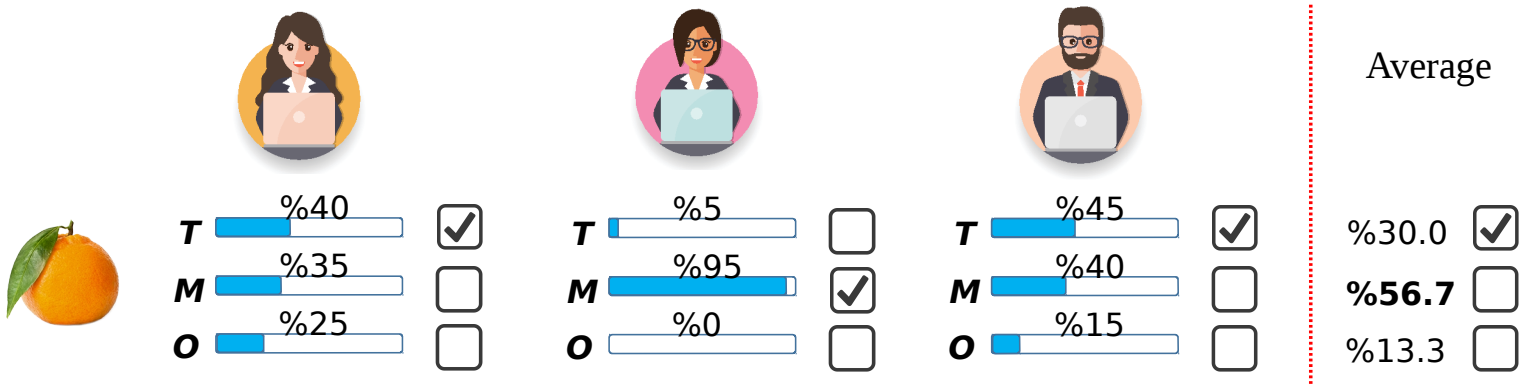
Estimating True Labels from Noisy Crowd Labels



Problem of aggregating crowd labels

- Crowd labels are often **noisy** and unreliable, since crowd workers are usually **inexpert** in the assigned tasks.
- To tackle this issue, each item is labeled multiple times by different workers. Our task is to **estimate the true labels** by **aggregating** these redundant crowd labels.

Single-option vs. multi-option crowd labels



Three crowd workers are asked to classify a figure as tangor (T), mandarin (M) or orange (O). Their single-option and multi-option crowd labels are shown with checked boxes and confidence bars respectively. The average score of multi-option labels correctly shows higher chance for mandarin, while the majority of single-option labels incorrectly suggests tangor as the truth.

Issue of single-option crowd labels

- Recent studies have shown that crowd workers cannot completely convey their **non-deterministic beliefs** with the single-option crowd labels!
- Flexible data collection mechanisms allow crowd workers to select **multiple options** for a question, or report their **non-deterministic confidence** level for the selected options.

Proposed Aggregation Models

Discriminative aggregation model

$CWMV_{\ell 1}$ objective function

$$\min_{\mathbf{w},\mathbf{y}_i \geq 0, \mathbf{1}^T\mathbf{y}_i=1} \quad \sum_{i=1}^{N} \|\mathbf{X}_i\mathbf{w} - \mathbf{y}_i\|_1 + \lambda_w \|\mathbf{w}\|_2^2$$

Joint Generative-Discriminative Aggregation Model

$DS\text{-}CWMV_{\ell 1}$ objective function

$$\min_{\mathbf{w},\mathbf{y}_i \geq 0, \mathbf{1}^T\mathbf{y}_i=1, \boldsymbol{\nu}_j} \quad \mathbf{KL}(p^{\boldsymbol{\nu}} \| p_0^{\boldsymbol{\nu}}) - \sum_{ijck} x_{ijk} y_{ic} \log(\nu_{jck}) + \sum_{i=1}^{N} \|\mathbf{X}_i\mathbf{w} - \mathbf{y}_i\|_1 + \lambda_w \|\mathbf{w}\|_2^2$$

Notation

$\mathbf{X}_i = [\mathbf{x}_{i1}, ..., \mathbf{x}_{iM}]^T$: crowd labels for the i-th item
$\mathbf{y}_i$: truth for the i-th item
$\boldsymbol{\nu}_j$: confusion matrix for the j-th worker (generative model parameter)
$\mathbf{w}$: workers reliability weights (discriminative model parameter)

Evaluation on Single-Option and Multi-Option Crowd Datasets

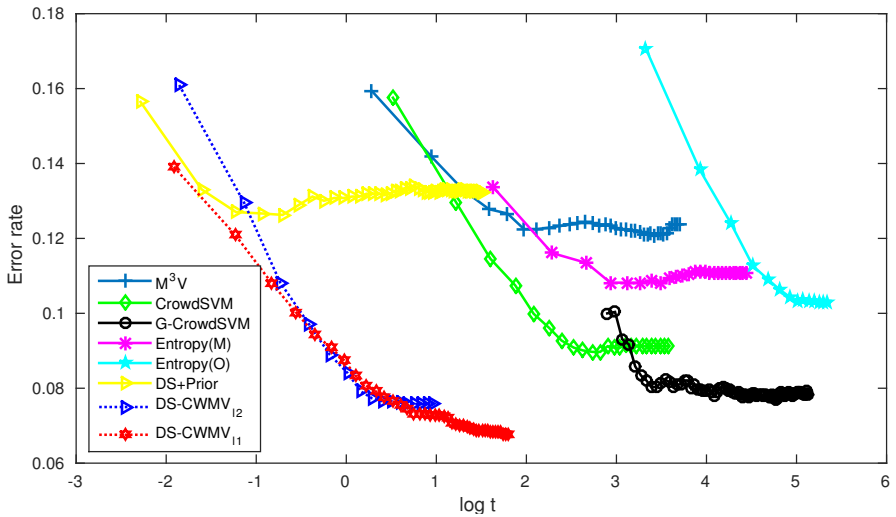
	Model	*Web Search*	*Age*	*RTE*	*Temp*	*Flowers*	Average
baselines	*MV*	26.90	34.88	10.31	6.39	22.00	28.83
	IWMV	15.04	34.53	8.12	5.84	19.00	17.09
	M^3V	12.74	33.33	7.88	6.06	13.50	15.43
	DS	16.92	39.62	7.25	5.84	13.00	18.69
	DS+Prior	13.26	34.53	[7.13]	5.84	13.50	15.80
	GLAD	19.30	35.73	**7.00**	[5.63]	13.50	16.20
	Entropy (M)	11.10	**31.14**	7.50	[5.63]	13.00	14.03
	Entropy (O)	10.40	37.32	-	-	-	17.76
	CrowdSVM	9.42	33.33	7.75	[5.63]	13.50	13.65
	G-CrowdSVM	7.99±0.26	32.98±0.36	7.67±0.19	5.71±0.33	[12.10±1.07]	12.78±0.31
ours	$CWMV_{\ell 2}$	10.89	34.43	7.25	[5.63]	16.00	14.65
	$CWMV_{\ell 1}$	10.70	34.23	7.50	[5.63]	13.00	14.43
	$DS\text{-}CWMV_{\ell 2}$	[7.58]	32.04	[7.13]	[5.63]	13.00	[12.32]
	$DS\text{-}CWMV_{\ell 1}$	**6.78**	[31.54]	**7.00**	**5.41**	**10.00**	**11.65**

Error rates (%) of crowdsourcing aggregation models on single-option crowdsourcing datasets.

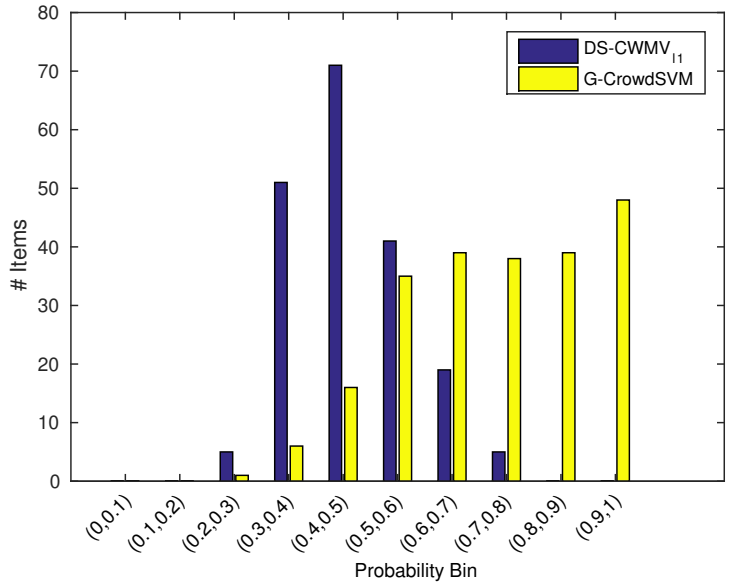
	Model	*Flag-Approval*	*Flag-Cumulative*	*Dog-Approval*	*Dog-Cumulative*	Average
baselines	*(soft) MV*	21.67	20.83	8.98	8.98	12.90
	(soft) DS	22.50	20.83	10.16	9.76	13.70
	(soft) DS+Prior	20.00	19.17	8.98	8.59	12.23
	(soft) Entropy (M)	17.50	16.67	12.89	12.89	14.23
ours	$CWMV_{\ell 2}$	16.67	16.67	8.98	8.59	11.30
	$CWMV_{\ell 1}$	[11.67]	[10.83]	[8.59]	[8.20]	[9.31]
	$DS\text{-}CWMV_{\ell 2}$	14.17	14.17	9.38	8.59	10.64
	$DS\text{-}CWMV_{\ell 1}$	**13.33**	**10.00**	**8.20**	**7.81**	**9.17**

Error rates (%) of crowdsourcing aggregation models applied on multi-option crowd datasets.

Evaluation of Running Speed and Reliability of Truths



Convergence comparison of aggregation models on *Web Search* dataset.



Histogram of the truths for mispredicted items. The results belongs to *DS-CWMV*$_{\ell 1}$ and *G-CrowdSVM* on *Web Search* dataset.