Accurate Crowd-Labeling using Item Response Theory

Faiza Khan Khattak Ansaf Salleb-Aouissi
Columbia University, New York

Anita Raja
The Cooper Union, New York.

Problem

- **Crowd-labeling:** Non-expert humans labeling a large dataset – tedious, time-consuming and expensive if accomplished by experts alone.
- Crowd-labelers are non-experts ➔ multiple labels per instance for quality assurance ➔ labels combined to get one final label.
- Many research papers [4,5,8] but still unresolved challenges.
- **Challenges:**
  - Getting accurate final label when crowd is of heterogeneous quality.
  - Identifying best ways to evaluate labelers.
  - Choosing per-class labeler ability or over all labeler ability.
  - Can quantifying prevalence of the class and/or clarity of a labeling task/question [9] improve accuracy?

Our Hypothesis

- Crowd-labeling ≈ Test taking.
- Item Response Theory (IRT) [12] used to model student ability (e.g., in GRE and GMAT.)
- IRT model:

$$P[A_i|\alpha_j,\delta_i,\beta_i] = [\text{logit}^{-1}(\delta_i(\alpha_j-\beta_i))]$$

A_i : Correct answer to question i,

α_j : ability of student j,

β_i : difficulty of question i, δ_i : clarity of question i

- IRT model is a compelling framework for crowd labeling.
- **Similarity:**
  - IRT model infers student and question related parameters, and probability of correctness of answers.
  - Crowd labeling process infers labeler and data instance related parameters, and probability of correctness of labels.
- **Difference:**
  - For IRT model correct answers are known.
  - For crowd labeling ground truth is to be inferred.

Our Approach

- A Bayesian approach to crowd-labeling inspired by IRT.
- New parameters and refined the usual IRT parameters to fit the crowd-labeling scenario.
- **Crowd Labeling Using Bayesian Statistics (CLUBS)**

$$P[c_k|y_{ij}=c_k,\gamma_{c_k},\beta_i,\delta_i,\pi_{c_k}^{(j)}] = [\text{logit}^{-1}(\delta_i(\gamma_{c_k}+\pi_c^{(j)}-\beta_i))]$$

where

c_k : class/category,

y_{ij} : Label provided by labeler j to instance i,

$\pi_{c_k}^{(j)}$: per-class ability of labeler j,

β_i : difficulty of instance i,

δ_i : clarity of question asked about instance i,

γ_{c_k} : prevalence of class c_k.

- **Parameters estimation:**
  - Since true labels are unknown, expert-labeled instances (ground truth) used for a small percentage of data (usually 0.1% -10%) for parameter estimation.
  - Estimated parameters used for aggregation of multiple crowd- labels for the rest of the dataset with no ground truth available.
  - Difficulty (β_i) and discrimination level (δ_i) of the instances without expert-labels are calculated as follows:

$$\hat{\beta}_i \sim \text{normal(mean.estimated-beta, sd.estimated-beta)}$$
$$\hat{\delta}_i \sim \text{normal(mean.estimated-delta, sd.estimated-delta)}$$

- **Label aggregation:**

Final label = $F_i = \text{sign}[\sum(P[c_k|y_{ij}=c,\gamma_{c_k},\hat{\beta}_i,\hat{\delta}_i,\pi_{c_k}^{(j)}] * y_{ij})]$

where

$$P[c_k|y_{ij}=c_k,\gamma_{c_k},\hat{\beta}_i,\hat{\delta}_i,\pi_{c_k}^{(j)}] = [\text{logit}^{-1}(\hat{\delta}_i(\gamma_{c_k}+\pi_{c_k}^{(j)}-\hat{\beta}_i))]$$

Experiments

- **Implementation**: Stan programming language [12].
- **State-of-the-art Methods:** We compared our method to Majority voting, Dawid and Skene [3], Expectation Maximization (EM), Karger's Iterative method (KOS) [7], Mean Field algorithm (MF) and Belief Propagation (BP) [10].
- **Datasets:** (a) Simulated dataset (b) Recognizing Textual Entailment (RTE)

Results

Table I. : Synthetic Data generation parameters and estimated parameters for the labelers. For the sake of presenting the labeler ability impact, the other parameters are kept fixed that instance difficulty $\beta \sim \mathcal{N}(0,2)$, instance question clarity $\delta \sim \mathcal{N}(0,0.75)$ and prevalence of class $\gamma = 0.5$.

		Dataset A	Dataset B
Class 1	% Correctness	(0.89, 0.90, 0.89, 0.95)	(0.98, 0.98, 0.85, 0.80)
	True Log-odds	(2.07, 2.18, 2.12, 3.01)	(3.89, 3.89,1.74,1.39)
	Estimated Log-odds	(1.42, 0.84, 1.42, 2.43)	(1.11, 0.64, 0.20, 0.20)
Class 2	% Correctness	(0.96, 0.83 0.98, 0.76)	(0.97, 0.89, 0.73, 0.74)
	True Log-odd	(3.18, 1.58, 4.23, 1.17)	(3.48, 2.09, 0.99, 1.05)
	Estimated Log-odds	(2.42, 1.00, 2.00, 2.00)	(1.19,1.50, 0.86,1.51)

Simulated data:

- Generated using fixed values of all the parameters except the labeler log-odds π (Table I).
- 5,000 instances, four crowd labels per instance, 20 expert-labeled instances.
- Accuracy (Table II.)

Method \ Dataset	A	B
MV	0.82	0.81
D &S	0.86	0.79
EM	0.77	0.76
BP (uniform prior)	0.82	0.83
MF (uniform prior)	0.77	0.76
KOS	0.75	0.76
Our approach	**0.89**	**0.89**

Table II. : Performance on Synthetic Data. Each dataset consists of 5,000 instances labeled by four labelers. Ground truth for 20 instances was taken as expert-labels.

RTE dataset:

- Five crowd labels per instance, 20 expert-labeled instances.
- Labeler error rate (Table III)
- Accuracy (Table IV.)

Table III. : Labeler performance for RTE Data.

% Error \ Labeler	L1	L2	L3	L4	L5
Overall	19.60	52.28	47.71	50.32	53.59
Class1	17.10	65.78	78.94	81.57	77.63
Class 2	22.07	38.96	16.88	19.48	29.87

Table IV. : Accuracy of final label for RTE Data.

Method \ Labelers	L1-L5	L1-L4	L1-L3	L1-L2	L2-L5	L2-L4	L2-L3
MV	0.55	0.52	0.61	0.57	0.47	0.50	0.50
D &S	0.41	0.46	0.47	**0.80**	0.46	0.47	0.48
EM	0.50	0.49	0.48	0.62	**0.50**	0.50	0.45
BP (uniform prior)	0.50	0.50	0.52	0.30	0.49	0.50	0.51
MF (uniform prior)	0.50	0.50	0.48	0.59	**0.50**	0.50	0.51
KOS	0.50	0.50	0.48	**0.80**	**0.50**	**0.51**	0.48
Our Approach	**0.65**	**0.70**	**0.73**	0.74	0.48	**0.51**	**0.54**

Conclusion

- Framework for crowd-labeling with detailed parameters.
- Results show better and stable performance when compared to state-of-the-art.
- Plan to make our approach more fine-grained by adding variability in labeler ability [1,2,13].

Contact

Faiza Khan Khattak. fk2224@columbia.edu
Ansaf Salleb-Aouissi. ansaf@cs.columbia.edu
Anita Raja. araja@cooper.edu

References

[1] Maarten A. S. Boksem, Theo F. Meijman, and Monicque M. Lorist. 2005. Effects of mental fatigue on attention: An ERP study. *Cognitive Brain Research* 25, 1 (Sept. 2005), 107–116. DOI:http://dx.doi.org/10.1016/j.cogbrainres.2005.04.011

[2] Arpád Csathó, Dimitri van der Linden, Istvan Hernádi, Peter Buzas, and Agnes Kalmar. Effects of mental fatigue on the capacity limits of visual attention. *Journal of Cognitive Psychology* 24, 5 (Aug. 2012), 511–524.

[3] A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. In *Applied Statistics*, Vol. 28. 20–28. Ofer Dekel and Ohad Shamir. 2009. Good learners for evil teachers. In *International Conference on Machine Learning ICML*. 30.

[4] Pinar Donmez and Jaime G. Carbonell. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)*. ACM, New York, NY, USA, 619–628.

[5] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL HLT, Atlanta, Georgia.*

[6] David Karger, Seewong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. In *Neural Information Processing Systems NIPS, Granada, Spain.*

[7] David R. Karger, Seweong Oh, and Devavrat Shah. 2014. Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems. *Operations Research* 62, 1 (2014), 1–24.

[8] Faiza Khan Khattak and Ansaf Salleb-Aouissi. 2013. Robust Crowd Labeling using Little Expertise. In *Sixteenth International Conference on Discovery Science, Singapore.*

[9] Aniket Kittur, H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proc. CHI 2008, ACM Pres.* 453–456.

[10] Qiang Liu, Jian Peng, and Alex Ihler. 2012. Variational Inference for Crowdsourcing. In *Advances in Neural Information Processing Systems NIPS,*

[11] P. Bartlett, F.c.n. Pereira, C.j.c. Burges, L. Bottou, and K.q.Weinberger (Eds.). 701–709. http://books.nips.cc/papers/files/nips25/NIPS2012 0328.pdf

[12] F. Lord. 1952. *A Theory of Test Scores*. Psychometric Monograph No. 7.
Stan Development Team. 2014. *Stan Modeling Language Users Guide and Reference Manual, Version 2.5.0.* http://mc-stan.org/

[13] Heikki Topi, Joseph S. Valacich, and Jeffrey A. Hoffer. 2005. The effects of task complexity and time availability limitations on human performance in database query tasks. *International Journal of Human-Computer Studies* 62, 3 (2005), 349–379.

[14] Jacob Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Neural Information Processing Systems NIPS.* 2035–2043.