Modeling Random Guessing and Task Difficulty for Truth Inference in Crowdsourcing

Extended Abstract

Yi Yang
Auckland University of Technology
Auckland, New Zealand
yi.yang@aut.ac.nz

Quan Bai
University of Tasmania
Hobart, Australia
quan.bai@utas.edu.au

Qing Liu
Data61, CSIRO
Hobart, Australia
q.liu@data61.csiro.au

ABSTRACT

This paper addresses the challenge of truth inference in crowdsourcing applications. We propose a generative method that jointly models tasks' difficulties, workers' abilities and guessing behavior to estimate the truths of crowdsourced tasks, which leads to a more accurate estimation on the workers' abilities and tasks' truths. Experiments demonstrate that the proposed method is more effective for estimating truths of crowdsourced tasks compared with the state-of-art methods.

KEYWORDS

Crowdsourcing; Truth Inference

ACM Reference Format:
Yi Yang, Quan Bai, and Qing Liu. 2019. Modeling Random Guessing and Task Difficulty for Truth Inference in Crowdsourcing. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

In a crowdsourcing platform, e.g. Amazon Mechanical Turk [1], a requester can post her tasks and obtain answers from the crowd workers. As the expertises and abilities of the workers are different, the collected answers to the same task are usually conflicting. Thus, an important task in crowdsourcing is to resolve the conflicts among the answers given by the crowd workers and discover the true answers (truths) for each task. Intuitively, we should trust answers from workers with high abilities. However, workers' abilities and tasks' truths are usually unknown *a priori*. Thus, truth inference [2, 3, 5, 8, 10–12] emerges and tackles this problem by jointly estimating workers' abilities and tasks' truths.

In this paper, we consider two important phenomenons in crowdsourcing applications for crowdsourcing truth inference. (1) The difficulties of crowdsourced tasks are usually different. A worker who can frequently answer easy questions correctly does not mean that her answers to the hard questions are also trustworthy. Thus, by modeling and estimating tasks' difficulties, the performance of truth inference is expected to be improved [6, 7, 9, 13]. (2) Most of the crowdsourced tasks are multi-choice tasks, i.e., each task has K mutual exclusive choices and there is only one true answer. As the workers in the crowdsourcing application are human, when a worker does not know the true answer of a task, she may choose to guess and submit a random answer.

Our contributions. Motivated by the two phenomenons described above, we propose a novel method, called ***C**rowdsourced **T**ruth **D**iscovery modeling **G**uessing and task **D**ifficulty* (CTDGD) , that infers multi-choice tasks' truths by jointly modeling tasks' difficulties and workers' abilities and guessing behavior. Specifically, the workers' abilities and answers and the tasks' true answers and difficulties are modeled as random variables in a probabilistic generative model. A worker's ability and the task's true answer and difficulty jointly determine if the worker knows the true answer of the task. If the worker does not know the truth, she submits a guessed answer from the available choices. By modeling guessing, the workers' abilities can be estimated without overestimation. By modeling tasks' difficulties, the truths of the hard tasks can be estimated more accurately. Experiments have been conducted on a real-world dataset and demonstrate that CTDGD outperforms the existing state-of-art crowdsourcing truth inference methods.

2 OUR METHOD

In this section, we present the CTDGD.

2.1 Answer Modeling

Suppose there are m workers $\{w_i\}_{i=0}^{m-1}$, and n tasks $\{t_j\}_{j=0}^{n-1}$. Each task has K mutual exclusive choices indexed from 1 to K. Each worker w_i can choose a choice as her answer x_{ij} for a task t_j. The goal of truth inference is to find the true answers $\{z_j\}$ for each task in $\{t_j\}$ from the observed answers $\{x_{ij}\}$. At the same time, the proposed CTDGD outputs the estimated workers' abilities $\{a_i\}$ and tasks' difficulties $\{d_j\}$.

We model each worker's ability a_i and each task's difficulty d_j as real numbers taken from $(-\infty, +\infty)$. Using the logistic function, the probability ϕ_{ij} of worker w_i knowing the true answer of t_j is

$$\phi_{ij} = \sigma(a_i - d_j) = \frac{1}{1 + \exp(-(a_i - d_j))} \tag{1}$$

where σ is the logistic function. From Equation (1) we can see that the probability of worker w_i knowing the truth of t_j is high if $a_i - d_j$ is large. Therefore, a worker is more likely to give a true answer to an easy task and less likely to answer a hard task correctly if her ability is smaller than the task's difficulty.

If the worker does not know the truth, she may guess and submit a random choice as her answer. Thus, the probability of an observed answer x_{ij} being the truth z_j is:

$$p(x_{ij} = k | z_j = k, d_j, a_i) = \phi_{ij} + (1 - \phi_{ij})\frac{1}{K} \tag{2}$$

[1] https://www.mturk.com/

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

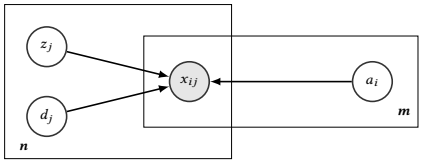

Figure 1: Graphical Model

We use the "one coin model" [13] to model the cases that a worker submits a wrong answer. Thus, for all $k' \neq k$, the probability of an observed answer x_{ij} being wrong is:

$$p(x_{ij} = k' | z_j = k, d_j, a_i) = (1 - \phi_{ij}) \frac{1}{K} \tag{3}$$

Combing Equations (2) and (3), the conditional probability of an observed worker's answer is:

$$p(x_{ij} | z_j, d_j, a_i) = \left(\phi_{ij} + (1 - \phi_{ij}) \frac{1}{K}\right)^{\delta_{ij}} \left((1 - \phi_{ij}) \frac{1}{K}\right)^{1 - \delta_{ij}} \tag{4}$$

where δ_{ij} denotes the Kronecker delta function.

2.2 Representation

CTDGD is a generative model. The worker's ability a_i, the task's difficulty d_j and truth z_j and the worker's answer x_{ij} are modeled as random variables. The relationships between these random variables are depicted in Figure 1. The generative processes of each random variable are described as follows.

The true answer is generated from a Categorical distribution: $p(z_j) = Cat(K, \boldsymbol{\alpha})$. The worker's answer is generated from a Categorical distribution with the p.m.f. defined in Equation (4). The task difficulty d_j is generated from a Normal distribution: $p(d_j) = \mathcal{N}(\mu_j, \sigma_j^2)$. The ability of a worker is generated from a Normal distribution: $p(a_i) = \mathcal{N}(\mu_i, \sigma_i^2)$. $\boldsymbol{\alpha}, \mu_j, \sigma_j^2, \mu_i$ and σ_i^2 are hyperparameters.

2.3 Inference

We use Expectation-Maximization (EM) algorithm to estimate the optimal values of $\{z_j\}$, $\{d_j\}$ and $\{a_i\}$. Specifically, we treat the true answers $Z = \{z_j\}$ as the latent variables, $\Theta = \{d_j\} \cap \{a_i\}$ as the model parameters , and $X = \{x_{ij}\}$ as the observations. The likelihood function is formulated in Equation (5).

$$L(\Theta; X, Z) = p(X, Z | \Theta) = \prod_j \left(p(z_j) \prod_i p(x_{ij} | z_j, d_j, a_i) \right) \tag{5}$$

EM algorithm finds the maximum likelihood of L and the optimal values of Z and Θ by iteratively performing an E-Step and a M-Step. In the E-Step, we compute $p_{jk}^{(t)}$, which is defined as the conditional probability $p(z_j = k | \Theta^{(t)}, X)$ at the current iteration t:

$$p_{jk}^{(t)} = \frac{p(z_j = k) \prod_i p(x_{ij} | z_j = k, d_j, a_i)}{\sum_{k'=1}^{K} p(z_j = k') \prod_i p(x_{ij} | z_j = k', d_j, a_i)} \tag{6}$$

In the M-step, we re-estimate the model parameters Θ at the next iteration $t + 1$ by maximizing an auxiliary function $Q(\Theta | \Theta^{(t)}) = E_{Z | \Theta^{(t)}, X} \left[\ln L(\Theta; X, Z) \right]$. There is no closed form to compute a_i and d_j directly to maximize Q. Therefore, we adopt gradient ascent to maximize Q, and the gradient of Q can be constructed by differentiating Q w.r.t. a_i and d_j:

$$\frac{\partial Q}{\partial a_i} = \sum_j \sum_{k=1}^{K} p_{jk}^{(t)} \left[\delta_{ij} \frac{K}{(K-1) + \frac{1}{\phi_{ij}}} - \phi_{ij} \right] \tag{7}$$

$$\frac{\partial Q}{\partial d_j} = - \sum_i \sum_{k=1}^{K} p_{jk}^{(t)} \left[\delta_{ij} \frac{K}{(K-1) + \frac{1}{\phi_{ij}}} - \phi_{ij} \right] \tag{8}$$

Given the above derivations, EM algorithm iteratively conducts the E-step and M-step until convergence. After the EM algorithm terminates, we can use the parameters in the last iteration as the estimated worker's ability and task's difficulty. At the same time, we can compute the estimated truth $\hat{z_j}$ by selecting the k^{th} choice that has the highest probability among $p_{jk}^{(t)}$, i.e., $\hat{z_j} = \arg\max_k \{p_{jk}^{(t)}\}$.

Method	Accuracy			
	Easy (1627) (Levels 1 - 7)	Medium (213) (Levels 8 - 9)	Hard (51) (Levels 10 - 12)	Overall (1891)
CTDGD	**96.25**	**75.14**	**64.71**	**93.02**
ZC [3]	94.53	66.2	52.94	90.22
GLAD [9]	94.28	66.2	50.98	89.95
DS [2]	95.02	66.66	50.95	90.64
LFC [8]	95.82	71.83	58.82	92.12
3Estimate [4]	95.88	71.83	60.5	92.23
TruthFinder [12]	94.96	68.07	47.06	90.69
Majority Voting	94.84	67.13	47.06	90.43

Table 1: Experimental Results

3 EXPERIMENTS

We conduct experiments on a real-world dataset, **Game** [1], to compare CTDGD with the state-of-art truth inference methods. The Game dataset contains 1908 unique questions with 12 difficulty levels. 1891 questions are answered by 37,332 workers with 214.658 answers. The performance is measured **accuracy**, which is defined as the percentage of the number of correctly inferred questions divided by the total number of questions.

Due to space limitation, we divide the questions into three categories, Easy, Medium and Hard, and present the results in Table 1. We list the number of questions in each level in the parentheses. From Table 1, we can observe that CTDGD has the best overall performance. For the easy tasks, we can see that all the methods have a very high accuracy, even majority voting can achieve over 90% accuracy. However, for the medium and hard level tasks, the accuracies of all the methods are dropped below 90%. This is because many workers cannot answer difficult questions correctly. Among all the methods, CTDGD has the best performance on medium and hards tasks, which demonstrates the superiority of CTDGD by jointly modeling tasks' difficulty and workers' guessing behavior.

4 CONCLUSION

In this paper, we propose ***C**rowdsourced **T**ruth **D**iscovery modeling **G**uessing and task **D**ifficulty* (CTDGD), which jointly models tasks' difficulties, workers' guessing behavior and abilities to estimate tasks' truths. Experiments on a real-world dataset demonstrate that CTDGD is more effective to estimate the truths of crowdsourced tasks than the state-of-art truth discovery methods, especially when the tasks are difficult.

REFERENCES

[1] Bahadir Ismail Aydin, Yavuz Selim Yilmaz, and Murat Demirbas. 2017. A crowdsourced "Who wants to be a millionaire?" player. *Concurrency and Computation: Practice and Experience* (2017), e4168.

[2] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.

[3] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 469–478.

[4] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. 2010. Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 131–140.

[5] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 1187–1198.

[6] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 745–754.

[7] Jermaine Marshall, Munira Syed, and Dong Wang. 2016. Hardness-aware truth discovery in social sensing applications. In *Distributed Computing in Sensor Systems (DCOSS), 2016 International Conference on*. IEEE, 143–152.

[8] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research* 11, Apr (2010), 1297–1322.

[9] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*. 2035–2043.

[10] Yi Yang, Quan Bai, and Qing Liu. 2018. On the Discovery of Continuous Truth: A Semi-supervised Approach with Partial Ground Truths. In *International Conference on Web Information Systems Engineering*. Springer, 424–438.

[11] Yi Yang, Quan Bai, and Qing Liu. 2019. A probabilistic model for truth discovery with object correlations. *Knowledge-Based Systems* 165 (2019), 360–373.

[12] Xiaoxin Yin, Jiawei Han, and S Yu Philip. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering* 20, 6 (2008), 796–808.

[13] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552.