A Reinforcement Learning Framework for Eliciting High Quality Information

Zehong Hu
Nanyang Technological University
huze0004@e.ntu.edu.sg

Yang Liu
Harvard University
yangl@seas.harvard.edu

Yitao Liang
University of California, Los Angeles
yliang@cs.ucla.edu

Jie Zhang
Nanyang Technological University
zhangj@ntu.edu.sg

Abstract

Peer prediction is a class of mechanisms designed to elicit high-quality information from strategic human agents when there is no ground-truth for contribution verification. Despite its elegant design, peer prediction is often found to have two shortcomings: (1) agents' cost/incentives for exerting effort to produce high-quality information is assumed to be known; (2) agents are modeled as fully rational. Both are too ideal in practice, preventing peer prediction from being applied in real-world problems. In this paper, we propose *Reinforcement Peer Prediction*, aiming to tackle these two limitations. It is the first model-free reinforcement learning (RL) framework, to our best knowledge, proposed in this domain. In our framework, a data requester dynamically learns agents' incentive level, using peer prediction scores, to maximize his own expected revenue. We leverage Bayesian inference to define unbiased estimators of the requestors' revenue and agents' states (accuracy levels). Those estimators form a core part in our RL process. Experiments show significant increment in the requestor's revenue under different agent models.

1 Introduction

Crowdsourcing rises as a promising method to collect a large amount of training data quickly under a limited budget. For example, a popular application of crowdsourcing is to generate labels for large scale datasets such as RTE [14] and ImageNet [4]. Notwithstanding its high efficiency, one salient concern about crowdsourcing is the quality of collected information, because it is often difficult or too costly to verify workers' contributions. This problem is called information elicitation without verification [15]. A class of incentive mechanisms, collectively called peer prediction, has been developed to solve this problem [10, 7, 18, 17, 11]. The core idea of peer prediction is quite simple and elegant – the mechanism designer financially incentivizes one worker via scoring his contribution by comparing it with those from his peers, and the payment rules are designed so that each worker reporting truthfully or reporting a high quality signal is a strict Bayesian Nash Equilibrium.

Effort-sensitive model is also adopted in peer prediction mechanisms [16, 2, 12, 8] to model agents' trade-off reasoning in contributing high-quality information. In these mechanisms, workers are incentivized to exert high efforts to generate high-quality answers. All the aforementioned peer prediction mechanisms focus on a one-shot interaction with agents, and one critical assumption is an explicitly-known worker model which includes workers' incentives to exert effort. Further it is also assumed workers are fully rational and only take the utility-maximizing strategy. Unfortunately, neither of above is true in practice. First of all, mostly likely workers' incentives, or cost, in exerting high effort can only be known after we, as the mechanism designers, interact with them. Secondly,

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

there is strong evidence that human workers are not fully rational [13]. And they are often observed to be deviating from equilibrium strategies in practice [9, 6].

To push peer prediction mechanisms towards the more practical ends, we propose a reinforcement learning framework, *Reinforcement Peer Prediction*, to interact with workers, so we will be able to (1) incentive workers to converge to a high effort exertion state (2) learn the optimal payment based on workers' contributions at each step. Nevertheless, there are two main challenges. Firstly, classic reinforcement learning focuses on the interaction between a single agent and its environment. We, instead, effectively consider a multi-agent setting, and immediately there is a game among workers, due to the peer prediction nature of our incentive structure. Therefore, the evolution of workers' state is a outcome of collective actions from all workers, as well as our environment. Secondly, no ground-truth answers are available for evaluating the reward, as often referred to in the reinforcement learning framework. Hence, we need to find a proper way to evaluate workers' contributions so that reward function based reinforcement learning algorithms can be applied.

The main contributions of this paper are as follows. (1) We propose the first model-free reinforcement peer prediction mechanism. Our mechanism combines the traditional peer prediction mechanisms with reinforcement learning to jointly incentive workers, as well as to learn and adjust the incentive level to offer. (2) Due to the missing of ground-truth, we adopt Bayesian inference to evaluate workers' contributions, to infer the reward following each action (offered incentive level). However, classic Bayesian inference algorithms for crowdsourcing (e.g. the Dawid-Skene estimator [3]) suffer from local optimum, causing the estimation of workers' contributions to be biased. This bias will be further amplified, due to the closed-loop nature of reinforcement learning. We derive the explicit posterior distribution of workers' contributions and employ Gibbs sampling for inference to eliminate the bias. (3) In our mechanism, the inferred contributions are corrupted by noise and we can only observe the last step worker state rather than the current one. Hence, in reinforcement learning, we use the online Gaussian process regression to learn the Q-function and replace the unknown current state with the couple of the last step state and incentive level. (4) We conduct empirical evaluation and show that our mechanism is robust, and is able to significantly increase data requester's revenue under different worker models, such as full rational and learning agent models.

2 Problem Formulation

Suppose in our system there are one data requester and N candidate workers denoted by $\mathcal{C} = \{1, \ldots, M\}$, where $M \geq 4$. At every step, the data requester assigns M binary answer tasks, with answer space $\{1, 2\}$, to workers. At step t, for task i, worker j's label can be written as $L_t(i, j)$, and correspondingly our mechanism needs to pay $P_t(i, j)$. Besides, we use $L_t(i, j) = 0$ to denote that task i is not assigned to worker j, and thus $P_t(i, j) = 0$. After collecting labels from workers, the data requester will aggregate labels via Bayesian inference [19], and the label accuracy can be written as A_t. Thus, the revenue of the data requester at step t can be computed as $r_t = F(A_t) - \eta \sum_{i=1}^{N} \sum_{j=1}^{M} P_{ij}$, where $F(\cdot)$ is a non-decreasing monotone function that maps accuracy into revenue. Intuitively the higher accuracy, the better. Meanwhile, the collected labels can only be used when their accuracy reaches a certain requirement. In this paper, we set $F(A_t) = A_t^{10}$. Note that our framework does not require any specific formulation of the F function. Suppose our mechanism goes for T steps. Our goal is to maximize the accumulative revenue, namely $R = \sum_{t=1}^{T} r_t$.

3 Reinforcement Peer Prediction

We present *Reinforcement Peer Prediction* in Figure 1. At step t, our mechanism decides the payment for workers as $P_t(i, j) = I_t \cdot p_t(i, j)$, where I_t denotes the incentive level learned and computed by our reinforcement learning algorithm. $p_t(i, j)$ denotes the output of the peer prediction mechanism. We use Bayesian inference to aggregate the collected labels. Since the ground-truth answers are unavailable, we cannot directly compute the reward, i.e. the accuracy A_t, following each action. Thus, we use the expected accuracy $\mathbb{E}A_t$ instead. It can be calculated as $\mathbb{E}A_t = \frac{1}{N} \sum_{i=1}^{N} \Pr(L_i = y_i)$, where L_i and y_i denote the aggregated and true label, respectively. Meanwhile, Bayesian inference also helps us output the confusion matrices of all workers and the distribution of task labels $[\Pr(l = 1), \Pr(l = 2)]$, with l denoting the ground-truth label. For worker j, his confusion

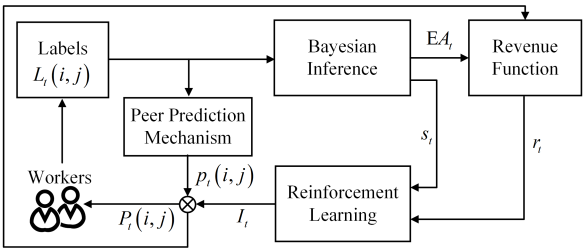


Figure 1: Illustration of our mechanism

matrix $C_j = [c_{jkg}]_{2\times 2}$ is a measurement of how much efforts he exerts, where c_{jkg} denotes the probability that worker j labels a task in class k as class g. Since the accuracy of labels are determined by the overall efforts of workers, how an individual worker performs is not the concern of our mechanism. Thus, we denote the state of the whole worker crowd s_t by workers' average probability of being correct, namely $s_t = \sum_{k\in\{1,2\}} \Pr(l=k) \cdot \frac{1}{M} \sum_{j=1}^{M} c_{jkk}$. After receiving the payment, workers may adjust their strategies in exerting efforts, which will lead to the change of s_t. However, when deciding the incentive level I_{t+1}, we only has the last step state s_t. In other words, the state observation has one-step delay, which also makes our reinforcement learning problem different from traditional ones.

Peer Prediction Mechanism: We adopt the multi-task mechanism proposed by [2]. For each task-worker pair (i,j), it selects a reference worker k. Suppose workers j and k have been assigned d other distinct tasks $\{j_1,\ldots,j_d\}$ and $\{k_1,\ldots,k_d\}$, respectively. Then, the payment $p_t(i,j) = 1[L(i,j) = L(i,k)] - \xi_j^d \cdot \xi_k^d - \bar{\xi}_j^d \cdot \bar{\xi}_k^d$, where $\xi_k^d = \sum_{g=1}^{d} 1(L(i_g,k)=1)/d$ and $\bar{\xi}_k^d = 1 - \xi_k^d$.

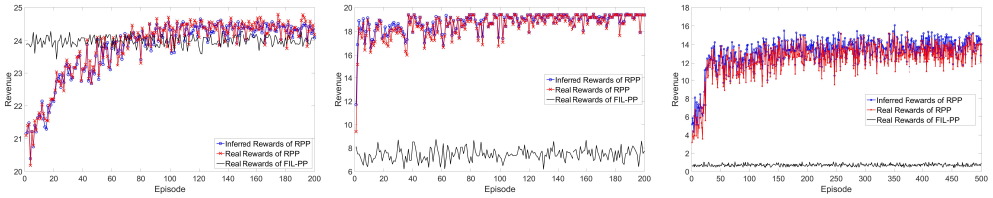
Bayesian Inference: Suppose the prior distributions that $c_{jk1} \sim \text{Beta}(\alpha_{jk1}^0, \alpha_{jk2}^0)$ and $\Pr(l=1) \sim \text{Beta}(\beta_1^0, \beta_2^0)$. Then, we can explicitly derive the posterior distribution of true labels as

$$P(\boldsymbol{y}|\boldsymbol{L}) \propto B(\boldsymbol{\beta}) \prod_{j=1}^{M} \prod_{k=1}^{K} B(\boldsymbol{\alpha}_{jk}) \ , \ \alpha_{jkg} = \sum_{i=1}^{N} \delta_{ijg}\xi_{ik} + \alpha_{jkg}^0, \beta_k = \sum_{i=1}^{N} \xi_{ik} + \beta_k^0 \quad (1)$$

where $B(\cdot)$ denotes the beta function, $\delta_{ijg} = 1(L(i,j)=g)$ and $\xi_{ik} = 1(y_i = k)$. According to Gibbs sampling, we can generate posterior samples via iteratively sampling $P(y_i|\boldsymbol{L}, \boldsymbol{y}_{j\neq i})$.

Reinforcement Learning: Recall that when computing the incentive level I_t for step t, the current state s_t cannot be observed. Thus, we define our incentive policy as $\pi(I_t|x_t)$, where $x_t = \langle s_{t-1}, I_{t-1}, t\rangle$. Then, the Q-function of our policy can be calculated as $Q(x_t, I_t) = \sum_{i=0}^{T-t} \gamma^i r_{t+i}$, where the discount factor γ is slightly smaller than 1 and $Q(x_{T+1}, I_{T+1}) = 0$. Both the state s_t and reward r_t are not accurately observed. Thus, we calculate the temporal difference of the Q-function as $Q(x_t, I_t) - \gamma Q(x_{t+1}, I_{t+1}) = r_t + N(x_t, x_{t+1})$, where the residual $N(x_t, x_{t+1})$ is assumed to be a Gaussian process. By applying the online Gaussian process regression algorithm [5], we can learn the Q-function effectively. Then, we decide the incentive level I_t for step t as $\arg\max Q(x_t, I_t)$ with probability $1-\epsilon$ and a random value with probability ϵ, which is the classic ϵ-greedy policy.

4 Experiments



(a) Rational Workers (b) Quantal Response Workers (c) MWUA Workers

Figure 2: Experiments on three popular worker models

Figure 2 show our experiment results on three popular worker models. Suppose there are four incentive levels, namely $I_t \in \{0.1, 1.0, 5.0, 10.0\}$. In practice, traditional peer prediction mechanisms are often applied with a fixed incentive level. Here, we set the incentive level as 1.0 and denote them by FIL-PP. By contrast, our reinforcement peer prediction (RPP) mechanism can dynamically learn the incentive level to maximize data requester's revenue. Besides, in our experiments, rational workers work with high efforts and report the true labels with probability 0.9 for any incentive level. Quantal response workers decide their strategy using the quantal response model, a classic bounded rationality model [9]. MWUA workers adapt their strategies via the MWUA model, a classic learning agent model [1]. From all the experiments, we can find that our mechanism is robust, and is able to significantly increase data requester's revenue, especially when workers are not fully rational.

References

[1] E. Chastain, A. Livnat, C. Papadimitriou, and U. Vazirani. Algorithms, games, and evolution. *PNAS*, 111(29):10620–10623, 2014.

[2] A. Dasgupta and A. Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proc. of WWW*, 2013.

[3] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*, 2009.

[5] Y. Engel, S. Mannor, and R. Meir. Reinforcement learning with gaussian processes. In *Proc. of ICML*, 2005.

[6] R. Jurca and B. Faltings. Robust incentive-compatible feedback payments. In *Agent-Mediated Electronic Commerce. Automated Negotiation and Strategy Design for Electronic Markets*, pages 204–218. Springer, 2007.

[7] R. Jurca, B. Faltings, et al. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34(1):209, 2009.

[8] Y. Liu and Y. Chen. Sequential peer prediction: Learning to elicit effort using posted prices. In *AAAI*, pages 607–613, 2017.

[9] R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

[10] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

[11] G. Radanovic and B. Faltings. A robust bayesian truth serum for non-binary signals. In *Proc. of AAAI*, 2013.

[12] V. Shnayder, A. Agarwal, R. Frongillo, and D. C. Parkes. Informed truthfulness in multi-task peer prediction. In *Proc. of ACM EC*, 2016.

[13] H. A. Simon. Rational decision making in business organizations. *The American economic review*, 69(4):493–513, 1979.

[14] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP*, 2008.

[15] B. Waggoner and Y. Chen. Output agreement mechanisms and common knowledge. In *Proc. of HCOMP*, 2014.

[16] J. Witkowski, Y. Bachrach, P. Key, and D. C. Parkes. Dwelling on the negative: Incentivizing effort in peer prediction. In *Proc. of HCOMP*, 2013.

[17] J. Witkowski and D. C. Parkes. Peer prediction without a common prior. In *Proc. of ACM EC*, 2012.

[18] J. Witkowski and D. C. Parkes. A robust bayesian truth serum for small populations. In *Proc. of AAAI*, 2012.

[19] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng. Truth inference in crowdsourcing: is the problem solved? *Proc. of the VLDB Endowment*, 10(5):541–552, 2017.