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 that help elicit high-quality information from strategic human agents when there is no ground-truth for verifying contributions. Despite its elegant design, peer prediction mechanisms often fail in practice, mostly due to two shortcomings: (1) agents' incentives for exerting effort are assumed to be known; (2) agents are modeled as being fully rational. In this paper, we propose the first reinforcement learning (RL) framework in this domain, *Reinforcement Peer Prediction*, to tackle these two limitations. In our framework, we develop a model-free RL algorithm for the data requester to dynamically adjust the incentive level to maximize his revenue, and to pay workers using peer prediction scoring functions. Experiments show significant improvement in data requestor's revenue under different agent models.

1 Introduction

Crowdsourcing rises as a promising inexpensive method to collect a large amount of training data quickly [14, 4]. Notwithstanding its high efficiency, one salient concern about crowdsourcing is the quality of the collected information, as it is often 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 workers according to the scoring of their contributions in comparison with their peers'. The payment rules are designed so that each worker reporting truthfully or reporting a high-quality signal is a strict Bayesian Nash Equilibrium.

Many peer prediction mechanisms adopt the effort-sensitive model to depict agents' trade-off reasoning in contributing high-quality information [16, 2, 12, 8]. In these mechanisms, workers are incentivized to exert high efforts to generate high-quality answers. One critical assumption in those mechanisms is an explicitly-known worker model which includes workers' incentives. Furthermore it is also assumed workers are fully rational, following utility-maximizing strategy. Unfortunately, neither is true in practice. Firstly, workers' incentives to exert high effort can most likely only be known after we, as the mechanism designers, interact with them. Secondly, there is strong evidence showing 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 being more practical, we propose a reinforcement learning framework, *Reinforcement Peer Prediction*, to interact with workers, so we will be able to

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

(1) incentive workers to converge to a high effort exertion state, and (2) learn the optimal payment based on workers' contributions at each step. Nevertheless, we face two main challenges. Firstly, classic reinforcement learning focuses on the interaction between a single agent and its environment. We, instead, need to effectively consider a multi-agent setting. Immediately a game is formed among workers, because our incentive strategy relies on the comparison between workers' answers and their peers'. Therefore, the evolution of workers' state is a outcome of collective actions from all workers, as well as our environment. Secondly, no ground-truth is available to evaluate either the state of the workers or the rewards in our setting, whereas it is taken as granted in most other reinforcement learning frameworks. Hence, we need to find a proper way to evaluate workers' contributions so that model-free RL algorithms which learn based on the reward signal 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 a peer prediction mechanism with reinforcement learning to jointly incentive workers and learn the optimal incentive level at each step. (2) Due to the lack of ground-truth, we adopt Bayesian inference to evaluate workers' contributions, and to infer the reward following each action (i.e. offered incentive level). We derive the explicit posterior distribution of workers' contributions and employ Gibbs sampling to eliminate its bias. (3) In our setting, the inferred contributions are corrupted by noise and only the most recent previous state rather than the current can be observed. We use the online Gaussian process regression to learn the Q-function and replace the unknown current state with the pair of the last observed state and incentive level. (4) We conduct empirical evaluation, and the results show that our mechanism is robust, and is able to significantly increase data requester's revenue under different worker models, such as fully rational, bounded rational and learning agent models.

2 Problem Formulation

Our proposed mechanism mainly works in the setting where one data requester, at every step, assigns M binary tasks with answer space $\{1, 2\}$ to $N \geq 4$ candidate workers. At step t, worker i's labels for the task j is denoted as $L_i^t(j)$, and correspondingly the payment that the mechanism pays is denoted as $P_i^t(j)$. Note we use $L_i^t(j) = 0$ to denote task j is not assigned to worker i at step t, and naturally under this case $P_i^t(j) = 0$. After every step, the data requester collects labels and aggregates them through Bayesian inference [19]. Assuming the aggregate reaches an accuracy A_t, the revenue for the data requester can then be computed as $r_t = F(A_t) - \eta \sum_{i=1}^{N} \sum_{j=1}^{M} P_i^t(j)$, where $F(\cdot)$ is a non-decreasing monotone function mapping accuracy to revenue and η is a tunable parameter balancing label quality and costs. Intuitively, $F(\cdot)$ needs to be non-deceasing as higher accuracy is preferred and also labels are only useful when their aggregate accuracy reaches a certain requirement. Note our framework is robust to any such $F(\cdot)$. For instance, it is set as $F(A_t) = A_t^{10}$ in our experiments. Our goal is to maximize the cumulative revenue $R = \sum_{t=1}^{T} \gamma^t r_t$, where γ is the discount rate and T is the ending time, by making wise choices of actions (i.e. deciding incentive levels) at each step.

3 Reinforcement Peer Prediction

We present *Reinforcement Peer Prediction* in Figure 1. Note at every step t, the payment to worker i for task j is factored as a function over two elements, namely $P_i^t(j) = I_t \cdot p_i^t(j)$, where $I_t \geq 0$ is the incentive level (scaling factor) learned and computed by our RL algorithm, and $p_i^t(j)$ is the output of our peer prediction scoring function. Since the ground-truth is not available, we cannot directly compute the reward (i.e accuracy A_t), following each action (i.e. deciding the offered incentive level). Thus, we introduce the expected accuracy $\mathbb{E}A_t$ as an unbiased estimator of the real accuracy A_t. It can be calculated as $\mathbb{E}A_t = \frac{1}{M} \sum_{j=1}^{M} \Pr(L^t(j) = y_j^t)$, where $L^t(j)$ and y_j^t are the aggregate and true labels at step t, respectively. Besides used to aggregate data, Bayesian inference is also used to generate 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 i, his confusion matrix $C_i = [c_{ikg}]_{2\times 2}$ is a measurement of how much efforts he exerts, where c_{ikg} denotes the probability that worker i labels a task in class k as class g. Since the accuracy of labels is determined by the overall efforts of workers, we denote the state of the whole worker crowd s_t by workers' average probability of being correct,

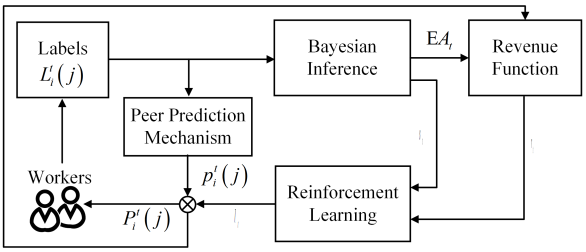


Figure 1: Illustration of our mechanism

namely $s_t = \frac{1}{N} \cdot \sum_{k=1}^{2} \Pr(l = k) \cdot \sum_{i=1}^{N} c_{ikk}$. After receiving the payment, workers may potentially adjust their strategies in exerting efforts (based on their rationality model), which will lead to the change of their state s_t. However, when deciding the next-step incentive level I_{t+1}, we can only estimate the last step state s_t. In other words, the state observation in our mechanism is not only imperfect, but also has one-step delay. This also makes our reinforcement learning problem different from traditional ones.

Peer Prediction Mechanism: We adopt the multi-task mechanism proposed by Dasgupta and Ghosh [2]. For each worker-task pair (i, j), it selects a reference worker k. Suppose workers i and k have been assigned d other distinct tasks $\{i_1, \ldots, i_d\}$ and $\{k_1, \ldots, k_d\}$, respectively. Then, the payment $p_i^t(j) = 1[L_i^t(j) = L_k^t(j)] - \xi_i^d \cdot \xi_k^d - \bar{\xi}_i^d \cdot \bar{\xi}_k^d$, where $\xi_i^d = \sum_{g=1}^{d} 1(L_k^t(i_g) = 1)/d$ and $\bar{\xi}_i^d = 1 - \xi_i^d$.

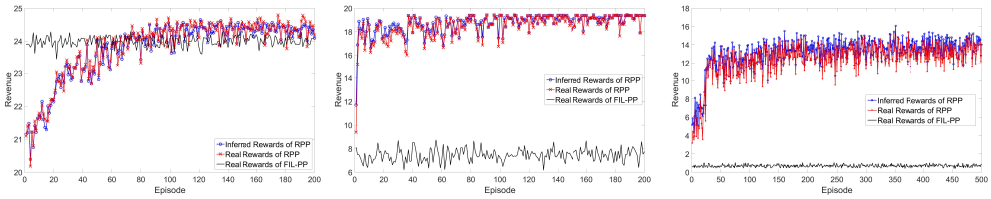
Bayesian Inference: Suppose the prior distributions are as follows: $c_{ik1} \sim \text{Beta}(\alpha_{ik1}^p, \alpha_{ik2}^p)$ and $\Pr(l = 1) \sim \text{Beta}(\beta_1^p, \beta_2^p)$. Then, we can explicitly derive the posterior distribution of true labels of assigned tasks, using collected labels from workers, as follows:

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

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

Reinforcement Learning: Recall that when computing the incentive level I_t for step t, the current state s_t cannot be observed. Because of it, we resort to the previous state and incentive level to define our policy $\pi(I_t|x_t)$, where $x_t = \langle s_{t-1}, I_{t-1}, t\rangle$. Then, the Q-function of our policy can be calculated as $Q(x_t, I_t) = \sum_{i=0}^{T-t} \gamma^i r_{t+i}$. Since both the state s_t and reward r_t can not be accurately observed, we have to approximate the temporal difference (TD) by assuming that the residual follows a Gaussian process: $Q(x_t, I_t) - \gamma Q(x_{t+1}, I_{t+1}) = r_t + N(x_t, x_{t+1})$, where $N(\cdot)$ is the residual. Then the Q-function can be learned effectively using the online Gaussian process regression algorithm [5]. Furthermore, we use the classic ϵ-greedy policy to make decisions at every step.

4 Experiments



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

The above figure shows 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 one-shot peer prediction mechanisms are often implemented with a fixed incentive level. Here, we set the incentive level as 1.0 and denote this mechanism with a fixed incentive level as FIL-PP. By contrast, our reinforcement peer prediction (RPP) mechanism can dynamically learn and adjust the incentive level to maximize data requester's revenue. Note, in our experiments, rational workers exert 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]. Based on all experiment results, we 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.