Identifying and Rewarding Subcrowds in Crowdsourcing

Siyuan Liu and **Xiuyi Fan** and **Chunyan Miao**[1]

Abstract. Identifying and rewarding truthful workers are key to the sustainability of crowdsourcing platforms. In this paper, we present a clustering based reward mechanism that rewards workers based on their truthfulness while accommodating differences in workers' preferences. Experimental results show that the proposed approach can effectively discover subcrowds under various conditions; and truthful workers are better rewarded than less truthful ones.

1 INTRODUCTION

Identifying and rewarding truthful workers are key to the sustainability of crowdsourcing platforms. However, in consensus tasks [1], workers may have an unknown number of different trustful answers. To accommodate this, we propose a partitional clustering technique to identify and reward *subcrowds*, a group of workers having similar preferences and giving similar answers to the consensus tasks. Unlike many other clustering algorithms which requires the prior knowledge of the number of clusters, our approach estimates the number of clusters. Thus, we assign each worker to a single cluster and reward the worker based on the distance to the cluster center. Experimental results show that the proposed clustering approach is able to identify subcrowds even with a significant amount of the population being untruthful. Results also show that the workers will receive more rewards if they provide more truthful answers.

2 IDENTIFYING AND REWARDING SUBCROWDS

Suppose a crowdsourced consensus task is composed of N questions. The answers from a worker w for these N questions is a vector/point $v_w = [a_w^1, a_w^2, \ldots, a_w^N]$ in an N dimension space. Thus, our goal is to classify these answer vectors into clusters.

Since the number of subcrowds is unknown, we need to firstly develop a clustering algorithm that estimates the number of clusters as well as partitioning the space into clusters. The developed algorithm is shown in Algorithm 1. We first randomly select a small subset V' from the set of collected answer vectors V as observation points in Line 1. Then for each vector v' in the subset, we calculate the distance between v' and any other vector v in V to create the distance histogram $hist$ in Line 5. We use the discrete metric and the L2 norm to measure distances for discrete and continuous answers, as given in Equation 1 and Equation 2, respectively.

$$dist(x, y) = |\{(x_i, y_i)|x_i \neq y_i, i = 1, 2, \ldots, N\}|. \quad (1)$$

$$dist(x, y) = (\sum_{i=1}^{N} |x_i - y_i|^2)^{\frac{1}{2}}. \quad (2)$$

[1] Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore, email: {syliu, xyfan, ascymiao}@ntu.edu.sg

Procedure: FindCenter(V)
Input : V, collected answer vectors;
Output : C, a set of initial centers;

```
1  Randomly select observation points V' ⊂ V;
2  C = ∅; hist = ∅;
3  foreach v' ∈ V' do
4      foreach v ∈ V do
5          hist[dist(v, v')] += 1;
6      foreach d ∈ hist do
7          if d is a local maximum in hist then
8              C_d = {v ∈ V | dist(v', v) == d};
9              C = C ∪ mean(C_d);
10 return C as initial centers.
```

Algorithm 1: Initial center estimation.

In Lines 6 and 7, we identify all local maxima in the histogram, as a local maximum indicates a dense area. In Lines 8 and 9, we identify all points in a dense area and set an initial center to be the center of this area. After we repeat the procedure for all the vectors in V', the cumulated set of initial centers are returned as C. Then we assign all points to their nearest centers in C to form clusters and move to the procedure of merging them, as follows.

For each cluster, s, with center c_s, we first find its radius r_s, defined as the distance from c_s to the farthest point in s. Then, for every two clusters s and s', if the distance between the two centers c_s and $c_{s'}$ are smaller than their radius, s and s' are then merged to form s_m. When there are no clusters to be merged, S will be returned as the resulting clusters.

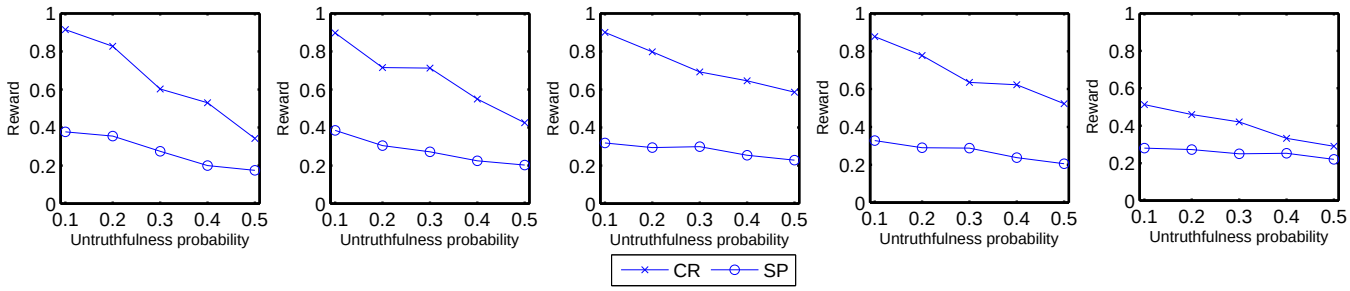
After clustering, we reward each worker based on its distance to its nearest cluster center. Namely, given a worker with answer vector v, let c_v be the cluster center that is closest to v, then the reward function R is:

$$R(v) = 1 - \frac{dist(v, c_v)}{N}. \quad (3)$$

The rewarding algorithm is based on the assumption that the distance $dist(v, c_v)$ increases as a worker's untruthfulness increases, which we believe is reasonable when the subcrowds share the same truthful answers and the workers behave consistently (within the subcrowd) upon providing answers for all the questions.

3 EXPERIMENTAL RESULTS

We use a discrete crowdsourcing dataset derived from [3] collected from Baidu Test to conduct experiments. Each test in the dataset is composed of 100 questions. Each question in the test contains 4 images and the task is to select the clearest one. The truthful answers can be classified into K types, where K is an integer in the range of



(a) $p = 0.1$ (b) $p = 0.2$ (c) $p = 0.3$ (d) $p = 0.4$ (e) $p = 0.5$

Figure 2. Comparison with the side-payment approach. CR is the proposed clustering reward; SP is the side-payment reward. $p = 0.1, \ldots, 0.5$ are population untruthfulness.

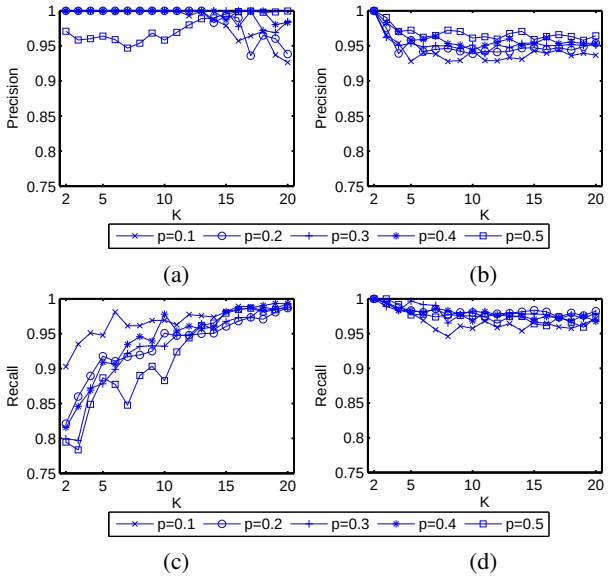


Figure 1. (a) Average precision for proposed approach; (b) Average precision for k-means; (c) Average recall for proposed approach; (d) Average recall for k-means

$[2, 20]$. For each type, we simulate 100 workers with untruthfulness p in $\{0.1, 0.2, \ldots, 0.9\}$ meaning that a worker has p chance to randomly select its answer[2]. Effectively, the dataset containing samples forming $2 - 20$ clusters with each cluster containing 100 vectors. The spread of a cluster is controlled by the untruthfulness p, e.g., for $p = 0.3$, for every answer that is in the sample, there is 30% chance that it is randomly selected. Each simulation is run 60 times and average results are presented.

We study the clustering accuracy (i.e., the accuracy in identifying subcrowds) using precision and recall[3]. To put our results into context, we compare our approach with the classic k-means with known k. Note that this gives k-means a strong edge as unlike feeding the correct k into k-means, our approach also estimates the number of clusters. Figure 1 presents the average precisions and recalls with worker untruthfulness in the range of $[0.1, 0.5]$.

As Figure 1 shows, the proposed approach achieves a higher precision result but lower recall result than k-means when the population untruthfulness is not greater than 0.5, suggesting that the proposed approach can achieve a high true positive. The recall increases with the number of ground truth clusters, suggesting that the false negative becomes smaller when there are more underlying clusters.

With the proposed clustering algorithm presenting promising results, we experiment its application on rewarding workers. We reused the dataset described before with 15 clusters and an additional worker changing his untruthfulness ap from 0.1 to 0.5 to study the reward achieved by the worker with different untruthfulness in various population truthfulness environments. Figure 2 plots the normalized reward result for the additional worker when the untruthfulness of the population also changes from 0.1 to 0.5.

In these figures, the x-axis is the worker's untruthfulness; and the y-axis is the normalized reward. We compare our results with the side-payment (SP) incentive mechanism [2]. SP rewards workers by comparing two randomly selected workers. Both workers are rewarded if their answers are identical; otherwise no worker is rewarded. From Figure 2, we can see that with our clustering based approach (i.e., CR), the worker with lower untruthfulness receives more reward, regardless how (un-)truthful the entire population is.

4 CONCLUSION

Developing mechanisms promoting worker truthfulness is a key problem in crowdsourcing. In this paper, we present a clustering based approach to identify subcrowds and reward workers based on the clustering result. The approach has the following advantages: (1) it identifies subcrowds even when there exist a large amount of untruthful answers; and (2) it rewards more to workers providing more truthful answers. In the future, we will continue improving the clustering techniques and conducting a more realistic testing.

ACKNOWLEDGEMENTS

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative.

REFERENCES

[1] D. C. Brabham, 'Crowdsourcing as a model for problem solving: An introduction and cases', *Convergence: The International Journal of Research into New Media Technologies*, **14**(1), 75–90, (2012).
[2] R. Jurca, *Truthful Reputation Mechanisms for Online Systems*, Ph.D. dissertation, EPFL, 2007.
[3] S. Liu, C. Miao, Y. Liu, H. Yu, J. Zhang, and C. Leung, 'An incentive mechanism to elicit truthful opinions for crowdsourced multiple choice consensus tasks', in *WI-IAT*, (2015).
[4] G. Sautter and K. Böhm, 'High-throughput crowdsourcing mechanisms for complex tasks', *Social Network Analysis and Mining*, **3**(4), 873–888, (2013).

[2] It is known that untruthful workers will mostly select random answers [4].
[3] Precision is defined as True Positive / (True Positive + False Positive); recall is defined as True Positive / (True Positive + False Negative).