OPTIMAL CROWDSOURCED CLASSIFICATION WITH A REJECT OPTION IN THE PRESENCE OF SPAMMERS

Qunwei Li, Pramod K. Varshney

Department of EECS, Syracuse University, Syracuse, NY 13244 USA
{qli33, varshney}@syr.edu

ABSTRACT

We explore the design of an effective crowdsourcing system for an M-ary classification task. Crowd workers complete simple binary microtasks whose results are aggregated to give the final decision. We consider the scenario where the workers have a reject option so that they are allowed to skip microtasks when they are unable to or choose not to respond to binary microtasks. We present an aggregation approach using a weighted majority voting rule, where each worker's response is assigned an optimized weight to maximize crowd's classification performance.

Index Terms— Classification, crowdsourcing, distributed inference, reject option, spammers

1. INTRODUCTION

Crowdsourcing provides a new framework to utilize distributed human wisdom to solve problems that machines cannot perform well, like handwriting recognition, paraphrase acquisition, audio transcription, and photo tagging [1–4]. In spite of the successful applications of crowdsourcing, the relatively low quality of output remains a key challenge [5–7].

Several methods have been proposed to deal with the aforementioned problems [8–15]. A crowdsourcing task is decomposed into microtasks that are easy for an individual to accomplish, and these microtasks could be as simple as binary distinctions [8]. A classification problem with crowdsourcing, where taxonomy and dichotomous keys are used to design binary questions, is considered in [9]. New aggregation rules that mitigate the unreliability of the crowd and improve the crowdsourcing system performance are investigated in [10, 11]. In our research group, we employed binary questions and studied the use of error-control codes and decoding algorithms to design crowdsourcing systems for reliable classification [9, 12]. A group control mechanism where the reputation of the workers is taken into consideration to partition the crowd accordingly into groups is presented in [13, 14]. Group control and majority voting techniques are compared in [15], which reports that majority voting is more cost-effective on less complex tasks. A weighted voting framework is developed in [16]. However, prior information of the individual workers is assumed known, which is unrealistic in most practical situations.

In past work on classification via crowdsourcing, crowd workers were required to provide a definitive yes/no response to binary microtasks. We consider the design of crowdsourcing systems where the workers are not forced to make a binary choice when they are unsure of their response and can choose not to respond. Crowd workers may be unable to answer questions for a variety of reasons such as lack of expertise. As an example, in mismatched speech transcription, i.e., transcription by crowd workers who do not know the language, workers may not be able to perceive the phonological dimensions they are tasked to differentiate [17]. We investigated the optimal aggregation rule where the workers have a reject option so that they are allowed to skip microtasks when they are unable to or choose not to respond [18, 19].

In this paper, we extend our work [18, 19] by further taking the spammers' effect on the system into consideration. We study the scenario where spammers also exist in the crowd, who participate in the task only to earn some free money without regard to the quality of their answers. The spammers submit answers with random guesses. We propose an optimal aggregation rule to combat the spammers' effect on system performance, which falls within the category of weighted majority voting methods, but where no prior individual information is needed. Simulation results show significant performance improvement by the proposed method.

2. CROWDSOURCING WITH A REJECT OPTION

Consider the situation where W workers take part in an M-ary object classification task. Each worker is asked N simple binary questions, termed as microtasks, which eventually lead to a classification decision among the M classes. We investigate independent microtask design and, therefore, we have $N = \lceil \log_2 M \rceil$ independent microtasks of equal difficulty. The workers submit results that are combined to give the final decision. Here, we consider the microtasks as simple binary questions and the worker's answer to a single microtask is conventionally represented by either "1" (Yes) or "0" (No) [9,20]. Thus, the wth worker's ordered answers to all the

microtasks form an N-bit word, which is denoted by $\mathbf{a}_w$. Let $\mathbf{a}_w(i)$, $i \in \{1, 2, \ldots, N\}$ represent the ith bit in this vector.

In our previous work [18, 19], we considered a more general problem setting where the worker has a reject option of skipping the microtasks. We denote this skipped answer as λ, whereas the "1/0" (Yes/No) answers are termed as definitive answers. Due to the variability of different workers' backgrounds, the probability of submitting definitive answers is different for different workers. Let $p_{w,i}$ represent the probability of the wth worker submitting λ for the ith microtask. Similarly, let $\rho_{w,i}$ be the probability that $\mathbf{a}_w(i)$, the ith answer of the wth worker, is correct given that a definitive answer has been submitted. Due to the variabilities and anonymity of workers, we study crowdsourcing performance when $p_{w,i}$ and $\rho_{w,i}$ are realizations of certain probability distributions, which are denoted by $F_P(p)$ and $F_\rho(\rho)$ respectively. The corresponding means are expressed as m and μ.

Let H_0 and H_1 denote the hypotheses where "0" or "1" is the true answer for a single microtask, respectively. For simplicity of performance analysis, H_0 and H_1 are assumed equiprobable for every microtask. The crowdsourcing task manager or a fusion center (FC) collects the N-bit words from W workers and performs fusion based on an aggregation rule.

In our previous work [18, 19], we proposed a novel weighted majority voting method for crowdsourced classification, which was derived by solving the following optimization problem

$$\begin{aligned} &\text{maximize} \;\; E_C[\mathbb{W}] \\ &\text{subject to} \;\; E_O[\mathbb{W}] = K \end{aligned} \tag{1}$$

where $E_C[\mathbb{W}]$ denotes the crowd's average weight contribution to the correct class and $E_O[\mathbb{W}]$ denotes the average weight contribution to all the possible classes that is constrained to remain a constant K. For ith bit, every worker's answer is assigned the derived optimal weight, and a decision is then obtained. We showed that this method significantly outperforms the widely-used simple majority voting procedure.

In this paper, we investigate the impact of spammers on system performance. The weight assignment scheme is developed by solving problem (1) as well.

3. OPTIMAL BEHAVIOR FOR THE MANAGER

In typical crowdsourcing setups, workers are simply paid in proportion to the number of tasks they complete [21]. Most likely, the spammers will complete all the microtasks with random guesses. A payment mechanism was proposed in the crowdsourcing system with a reject option to incentivize the crowd, where responses with even the slightest error are associated with minimum payment possible [21]. This mechanism promotes skipping of all the microtasks by the spammers. Therefore, we assume that M_A spammers complete all the microtasks and the rest of the M_0 spammers skip all the microtasks, making a total of M spammers in the crowd of size W. To combat the spammers' effect on the system performance, we develop the aggregation rule on the manager's side with a new weight assignment scheme to maximize the weight assigned to the correct class.

Proposition 1. *To maximize the average weight assigned to the correct classification element, the weight for the wth worker's answer is given by*

$$W_w = \left[(W - M)\,\mu^n + \frac{M_A}{2^N (1-m)^N} \delta(n - N) \right]^{-1}, \tag{2}$$

where n is the number of definitive answers that the wth worker submits, and $\delta(\cdot)$ is the Dirac delta function.

Proof. When there are M spammers in the crowd with M_0 skipping and M_A completing all the questions, the expected weight contributed to the correct class is given by

$$\begin{aligned} E_C[\mathbb{W}] &= \sum_{w=1}^{W-M} E_{p,\rho}\left[\sum_{n=0}^{N} W_w \rho(n) P_\lambda(n)\right] + \sum_{w=1}^{M_0} W_w(n=0) \\ &\quad + \sum_{w=1}^{M_A} \frac{1}{2^N} W_w(n=N) \\ &= \sum_{n=0}^{N} (W-M) W_w \mu^n \binom{N}{n} (1-m)^n m^{N-n} \\ &\quad + \sum_{n=0}^{N} M_0 W_w \delta(n) + \sum_{n=0}^{N} \frac{M_A}{2^N} W_w \delta(n-N) \\ &= \sum_{n=0}^{N} (W-M) W_w \mu^n \mathbb{P}(n) + \sum_{n=0}^{N} \frac{M_0}{\mathbb{P}(0)} W_w \mathbb{P}(n) \delta(n) \\ &\quad + \sum_{n=0}^{N} \frac{M_A}{2^N \mathbb{P}(N)} W_w \mathbb{P}(n) \delta(n-N) \\ &= \sum_{n=0}^{N} W_w S(n) \mathbb{P}(n), \end{aligned} \tag{3}$$

where $\mathbb{P}(n) = \binom{N}{n}(1-m)^n m^{N-n}$, and

$$S(n) = (W-M)\mu^n + \frac{M_0}{m^N}\delta(n) + \frac{M_A}{2^N(1-m)^N}\delta(n-N).$$

Note that $\sum_{n=0}^{N} \mathbb{P}(n) = 1$, and then (3) is upper-bounded using Cauchy-Schwarz inequality as follows:

$$\begin{aligned} E_C[\mathbb{W}] &= \sum_{n=0}^{N} W_w S(n) \mathbb{P}(n) \\ &\le \sqrt{\sum_{n=0}^{N} (W_w S(n))^2 \mathbb{P}(n)} \sqrt{\sum_{n=0}^{N} \mathbb{P}(n)} = \alpha. \end{aligned} \tag{4}$$

Also note that equality holds in (4) only if

$$W_w S(n)\sqrt{\mathbb{P}(n)} = \alpha\sqrt{\mathbb{P}(n)}$$

where α is a positive constant, and $W_w S(n) = \alpha$

Therefore, the optimal behavior for the manager in terms of the weight assignment is obtained as

$$W_w = \left[(W-M)\,\mu^n + \frac{M_0}{m^N}\delta\,(n) + \frac{M_A}{2^N(1-m)^N}\delta\,(n-N)\right]^{-1}.$$

Note that if a worker submits no definitive answers, i.e. $n = 0$, the corresponding weight assigned is $(W - M + \frac{M_0}{m^N})^{-1}$. However, since this worker skips all the questions, his/her decision for a certain question is not taken into consideration at the fusion center and, without loss of generality, the corresponding weight can be set equal to zero. Therefore, the weight assignment for the scheme can be expressed as

$$W_w = \left[(W-M)\,\mu^n + \frac{M_A}{2^N(1-m)^N}\delta\,(n-N)\right]^{-1}.$$

Compared to the weight assignment for an honest crowd [18], the derived scheme differs in terms of the weight assigned to the workers who complete all the microtasks. If the spammers skip all the microtasks, the weight assignment scheme remains the same, which is intuitively true as no random guesses are received by the manager from the spammers and the crowd can be considered as honest as well. Otherwise, the weight assignment scheme differs from the scheme given in [18].

3.1. Parameter Estimation

In order to act optimally, the manager has to estimate several parameters before the weight assignment can be adopted. Specifically, one has to estimate μ, m, M_A, M_0 before he/she can proceed with the optimal weight assignment. We can employ either the *Training-based* or *Majority-voting based* method to estimate μ as stated in our previous work [18]. Calculating the ratio of the sum of skipped questions over all the questions attempted by the crowd gives the estimated m. Based on the analysis in previous sections, the answers with all questions completed or skipped should be discarded for estimation.

We hereby jointly address the estimation of M_0 and M_A by using the maximum likelihood estimation (MLE) method. First, as we employ G gold standard questions, a worker has to respond to $N + G$ microtasks. Let W_{N+G} denote the number of workers submitting $N + G$ definitive answers, and W_0 denote the number of workers skipping all the microtasks. Given the numbers of spammers respectively completing and skipping all the microtasks, M_A and M_0, the joint probability distribution function of W_{N+G} and W_0, $f(W_{N+G}, W_0|M_A, M_0)$, is expressed in (6), where $\hat{m}$ is the estimated m.

Therefore, by the MLE method, the estimates of M_A and M_0, which are denoted by $\hat{M}_A$ and $\hat{M}_0$ respectively, can be obtained as

$$\left\{\hat{M}_A, \hat{M}_0\right\} = \underset{\{M_A, M_0\}\geq 0}{\arg\max}\; f(W_{N+G}, W_0|M_A, M_0). \quad (5)$$

Once the manager has the estimation results $\hat{\mu}$, $\hat{m}$, $\hat{M}_A$, and $\hat{M}_0$, he/she can optimally assign the weight to the workers' answers for aggregation.

3.2. Performance Analysis

In this section, we characterize the performance of such a crowdsourcing classification framework, where the task manager behaves optimally, in terms of the probability of correct classification P_c. Note that we have an overall correct classification only when all the bits are classified correctly.

Proposition 2. *The probability of correct classification P_c in the crowdsourcing system is*

$$P_c = \Big[\frac{1}{2} + \frac{1}{2}\sum_{S}\binom{W}{\mathbb{Q}}\left(F(\mathbb{Q}) - F'(\mathbb{Q})\right) + \frac{1}{4}\sum_{S'}\binom{W}{\mathbb{Q}}\left(F(\mathbb{Q}) - F'(\mathbb{Q})\right)\Big]^N \quad (7)$$

with

$$F(\mathbb{Q}) = m^{q_0}\prod_{n=1}^{N}(1-\mu)^{q_{-n}}\mu^{q_n}\left(C_{N-1}^{n-1}(1-m)^n m^{N-n}\right)^{q_{-n}+q_n}$$

and

$$F'(\mathbb{Q}) = m^{q_0}\prod_{n=1}^{N}(1-\mu)^{q_n}\mu^{q_{-n}}\left(C_{N-1}^{n-1}(1-m)^n m^{N-n}\right)^{q_{-n}+q_n}$$

where

$$\mathbb{Q} = \{(q_{-N}, q_{-N+1}, \ldots q_N, M_A', M_A'') : \sum_{n=-N}^{N} q_n = W - M_A - M_0, M_A' + M_A'' = M_A\},$$

with natural numbers q_n, M_A', and M_A'',

$$S = \left\{\mathbb{Q}: \sum_{n=1}^{N}\frac{q_n - q_{-n}}{(W-M)\mu^n} + (M_A' - M_A'')\frac{2^N(1-m)^N}{M_A} > 0\right\},$$

$$S' = \left\{\mathbb{Q}: \sum_{n=1}^{N}\frac{q_n - q_{-n}}{(W-M)\mu^n} + (M_A' - M_A'')\frac{2^N(1-m)^N}{M_A} = 0\right\},$$

and $\binom{W}{\mathbb{Q}} = \frac{W!}{\prod_{n=-N}^{N} q_n!}$.

$$f(W_{N+G}, W_0|M_A, M_0) = \binom{W_0 - M_0}{W - M_0 - M_A} (\hat{m}^{N+G})^{W_0 - M_0} (1 - \hat{m}^{N+G})^{W - W_0 - M_A} \cdot \binom{W_{N+G} - M_A}{W - W_0 - M_A} (1 - \hat{m})^{(N+G)(W_{N+G} - M_A)} \left(1 - (1 - \hat{m})^{N+G}\right)^{W - W_{N+G} - W_0} \quad (6)$$

Proof. Due to the space limit, we only give the result here. The proof will be given in the extended version of the paper and a similar proof can be found in our previous paper [18].

3.3. Simulation Results

In this section, we present the simulation results to illustrate the performance of the proposed schemes. $W = 50$ workers participate in a crowdsourcing task with $N = 3$ microtasks and $G = 3$ gold standard questions. $F_P(p)$ is chosen as a uniform distribution $U(0, 1)$, and let $F_\rho(\rho)$ be a uniform distribution expressed as $U(x, 1)$ with $0 \leq x \leq 1$, and thus we can have μ varying from 0.5 to 1.

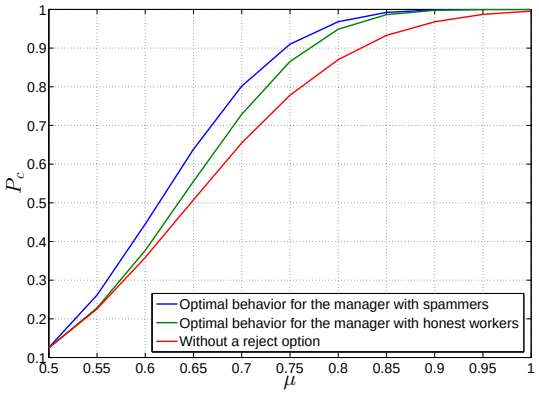

Fig. 1. Performance comparison with spammers.

We present the performance comparison with spammers in Fig. 1, where the quality of the crowd μ varies. We plot the performance of three different weight assignment methods. The first one is what we derived in this section, which is referred to as the optimal behavior for the manager with spammers. The second is the one that we derived in [18], which is given by $W_w = \mu^{-n}$. Since we do not assume the knowledge of prior information regarding individuals, the existing weighted majority voting methods fail to work in this setting. Thus, we choose the conventional simple majority voting without a reject option for comparison. For illustration, there are 14 spammers in a crowd of 50 workers, and we have 7 spammers completing all the questions and the other 7 skipping all the questions. When $\mu = 0.5$, the workers are making random guesses even if they believe that they are able to respond with definitive answers. In such a case, the choice of weight assignment schemes does not make a difference, and therefore, the three curves merge at this point. The method with optimal behavior for the manager with spammers outperforms the other two, while the simple majority voting without a reject option performs the worst.

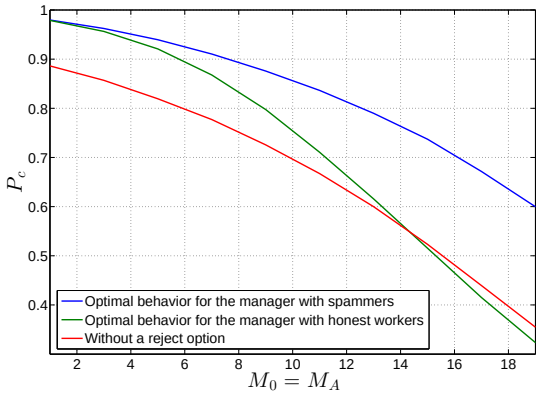

Fig. 2. Performance comparison with various spammers.

In Fig. 2, we plot the performance comparison when the number of spammers changes. We set that $M_0 = M_A$, and μ is fixed at 0.75. As we can observe, the method with optimal behavior for the manager with spammers yields the best performance. When the number of spammers is small, the simple majority voting method is outperformed by the one with optimal behavior for the manager with honest workers. However, this is not the case when the number of spammers is large. The reason is that with honest workers, the manager assigns a greater weight to the worker with a larger number of definitive answers. In the regime where M_A is large, which means the number of spammers completing all the questions is large, the impact from the spammers is much more severe on the performance with such a weight assignment scheme. Thus, the corresponding performance degrades significantly.

4. CONCLUSION

We have studied a novel framework of crowdsourcing system for classification, where an individual worker has the reject option and can skip a microtask if he/she has no definitive answer. We investigated the impact of the spammers in the crowd on the crowdsourcing system performance. We derived the optimal strategy for the manager, where an optimal weighted aggregation rule for the crowdsourcing was proposed to combat the spammers' influence.

5. REFERENCES

[1] P. Paritosh, P. Ipeirotis, M. Cooper, and S. Suri, "The computer is the new sewing machine: Benefits and perils of crowdsourcing," in *Proc. 20th Int. Conf. World Wide Web (WWW'11)*, Mar.–Apr. 2011, pp. 325–326.

[2] E. Kamar, S. Hacker, and E. Horvitz, "Combining human and machine intelligence in large-scale crowdsourcing," in *Proc. 11th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Jun. 2012, pp. 467–474.

[3] S. Burrows, M. Potthast, and B. Stein, "Paraphrase acquisition via crowdsourcing and machine learning," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 3, p. 43, Jul. 2013.

[4] J. Fan, M. Zhang, S. Kok, M. Lu, and B. C. Ooi, "CrowdOp: Query optimization for declarative crowdsourcing systems," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2078–2092, Aug. 2015.

[5] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on Amazon Mechanical Turk," in *Proc. ACM SIGKDD Workshop Human Comput. (HCOMP'10)*, Jul. 2010, pp. 64–67.

[6] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," *IEEE Internet Comput.*, no. 2, pp. 76–81, Mar. 2013.

[7] K. Mo, E. Zhong, and Q. Yang, "Cross-task crowdsourcing," in *Proc. ACM Int. Conf. Knowl Discovery Data Mining*, Aug. 2013, pp. 677–685.

[8] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Advances in Neural Information Processing Systems (NIPS)*, Dec. 2011, pp. 1953–1961.

[9] A. Vempaty, L. R. Varshney, and P. K. Varshney, "Reliable crowdsourcing for multi-class labeling using coding theory," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 667–679, Aug. 2014.

[10] D. Yue, G. Yu, D. Shen, and X. Yu, "A weighted aggregation rule in crowdsourcing systems for high result accuracy," in *Proc. IEEE 12th Int. Conf. Depend., Auton. Secure Comput. (DASC)*, Aug. 2014, pp. 265–270.

[11] D. Sanchez-Charles, J. Nin, M. Sole, and V. Muntes-Mulero, "Worker ranking determination in crowdsourcing platforms using aggregation functions," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ)*, Jul. 2014, pp. 1801–1808.

[12] L. R. Varshney, A. Vempaty, and P. K. Varshney, "Assuring privacy and reliability in crowdsourcing with coding," in *Proc. 2014 Inf. Theory Appl. Workshop*, Feb. 2014.

[13] A. J. Quinn and B. B. Bederson, "Human computation: a survey and taxonomy of a growing field," in *Proc. 2011 Annu. Conf. Hum. Factors Comput. Syst. (CHI 2011)*, May 2011, pp. 1403–1412.

[14] Y. Zhang and M. van der Schaar, "Reputation-based incentive protocols in crowdsourcing applications," in *Proc. 31st IEEE Conf. Computer Commun. (INFOCOM 2012)*, Mar. 2012, pp. 2140–2148.

[15] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms," *Math. Comput. Model.*, vol. 57, no. 11, pp. 2918–2932, Jul. 2013.

[16] L. I. Kuncheva and J. J. Rodríguez, "A weighted voting framework for classifiers ensembles," *Knowledge and Information Systems*, vol. 38, no. 2, pp. 259–275, Feb 2014.

[17] P. Jyothi and M. Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," in *Proc. 29th AAAI Conf. Artificial Intelligence (AAAI'15)*, Nov. 2015.

[18] Q. Li, A. Vempaty, L. R. Varshney, and P. K. Varshney, "Multi-object classification via crowdsourcing with a reject option," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 1068–1081, Feb 2017.

[19] Q. Li and P. K. Varshney, "Does confidence reporting from the crowd benefit crowdsourcing performance?" in *Proceedings of the 2nd International Workshop on Social Sensing*, ser. SocialSens'17, Apr. 2017.

[20] J. Rocker, C. M. Yauch, S. Yenduri, L. Perkins, and F. Zand, "Paper-based dichotomous key to computer based application for biological indentification," *J. Comput. Sci. Coll.*, vol. 22, no. 5, pp. 30–38, May 2007.

[21] N. B. Shah and D. Zhou, "Double or nothing: Multiplicative incentive mechanisms for crowdsourcing," in *Advances in Neural Information Processing Systems (NIPS)*, Dec. 2015, pp. 1–9.