Human Computation and Crowdsourcing: Works in Progress Abstracts:
An Adjunct to the Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing

Cheaper and Better: Selecting Good Workers for Crowdsourcing

Hongwei Li
Department of Statistics, UC Berkeley
hwli@stat.berkeley.edu

Qiang Liu
Computer Science, Dartmouth College
qliu@cs.dartmouth.edu

Abstract

Crowdsourcing provides a popular paradigm for data collection at scale. We study the problem of selecting subsets of workers from a given worker pool to maximize the accuracy under a budget constraint. One natural question is whether we should hire as many workers as the budget allows, or restrict on a small number of top-quality workers. By theoretically analyzing the error rate of a typical setting in crowdsourcing, we frame the worker selection problem into a combinatorial optimization problem and propose an algorithm to solve it efficiently. Empirical results on both simulated and real-world datasets show that our algorithm is able to select a small number of high-quality workers, and performs as good as, sometimes even better than, the much larger crowds as the budget allows. This is a short version of our full length paper (Li and Liu 2015) available at http://arxiv.org/abs/1502.00725.

Introduction

The micro-task crowdsourcing platforms, such as Amazon Mechanical Turk and crowdflower, provide a powerful tool for collecting large amounts of human-labeled data at a relatively low cost. However, due to the unreliability of the (often anonymous) crowd workers, a major challenge is to estimate the workers' reliabilities and aggregate their labels accordingly. A large body of work has been developed to address this problem; see (Dawid and Skene 1979; Karger, Oh, and Shah 2011; Liu, Peng, and Ihler 2012; Liu, Ihler, and Steyvers 2013) and references therein.

This work is motivated by a natural question: do more crowd workers necessarily yield better aggregated results than less workers? The idea of *wisdom of crowds* seems to suggest a confirmative answer, since "*larger crowds* should be *wiser*". However, in practice, because the workers' prediction model and reliabilities are never known perfectly, we run the risk of adding noisy information from the unreliable workers as we include all the workers. In fact, a recent empirical study (Mannes, Soll, and Larrick 2013) showed that the aggregated results of a small number of (3 to 6) high-quality workers are often more accurate than those of much larger crowds (a.k.a. *the wisdom of small crowds*). In this work, we study this phenomenon by formulating a worker selection problem under a budget constraint. We frame the problem into a combinatorial optimization that minimizes an upper bound of the error rate, and deriving a globally optimal algorithm that selects a group of top-ranked workers that optimize the upper bound of the error rate. We demonstrate the efficiency of our algorithm by comprehensive experiments on a number of real-world datasets.

Worker Selection in Crowdsourcing

Assume there are M crowd workers and N items (or questions) each with labels from L classes. We use $[M]$ to denote the set of first M integers. We assume each item j is associated with an unknown true label $y_j \in [L]$, $j \in [N]$. When item j is assigned to worker i for labeling, we get a possibly inaccuracy answer from the worker, which we denote by $Z_{ij} \in [L]$. The workers have different reliabilities, which can be characterized by their probability w_i of giving correct answers, that is, $w_i = \mathbb{P}(Z_{ij} = y_j)$. To access the reliability w_i, we assume that we have n *control* (or *gold standard*) questions with known true labels; this allows us to estimate w_i by $\hat{w}_i = c_i/n$, where c_i is the number of questions answered correctly by worker j; further, the uncertainty of $\hat{w}_i$ can be measured by the empirical variance $\hat{\text{var}}(\hat{w}_i) = \hat{w}_i(1-\hat{w}_i)/(n-1)$.

Given the labels Z_{ij} from a set S of workers, the true labels y_j can be estimated by the linear-weighted majority voting method as discussed in Li, Yu, and Zhou 2013,

$$\hat{y}_j = \underset{k\in[L]}{\operatorname{argmax}} \sum_{i\in S} (L\hat{w}_i - 1) \cdot \mathbb{I}\left(Z_{ij} = k\right), \qquad (1)$$

where the importance of the workers' answers are weighted according to their reliabilities $\hat{w}_i$. Other advanced label aggregation methods share a similar weighted majority form, excepting using a different weight function. For example, EM can be treated as using a log-odd weight $\log((L-1)\hat{w}_i/(1-\hat{w}_i))$. We prefer WMV-linear because it is more stable and simpler for theoretical analysis (Li, Yu, and Zhou 2013).

For the worker selection problem, we are interested in finding the optimal worker subset S of size less than a budget K out of the worker pool $\Omega = [M]$ to minimize the prediction error rate, that is,

$$\min_{S\subset\Omega}\left\{e(S)\equiv\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}(\hat{y}_j\neq y_j)\right\},\quad s.t.\quad |S|\leq K.$$

It is intractable to calculate the exact error rate. However, our full paper (Li and Liu 2015) shows that error rate is bounded by a non-increasing function of a factor which can be unbiasly estimated by

$$\hat{F}(S)=\frac{1}{\sqrt{|S|}}\sum_{i\in S}\left[(L\hat{w}_i-1)^2-L^2\hat{\mathrm{var}}(\hat{w}_i)\right],\qquad (2)$$

where the first term $(L\hat{w}_i-1)^2$ counts the reliability of the worker i, and the second term $\hat{\mathrm{var}}(\hat{w}_i)$ represents the empirical variance of $\hat{w}_i$, down-weighting these $\hat{w}_i$ with large estimation uncertainty. This allows us to reframe the selection problem into a much easier form,

$$\min_{S\subset\Omega}\hat{F}(S),\qquad s.t.\quad |S|\leq K.\qquad (3)$$

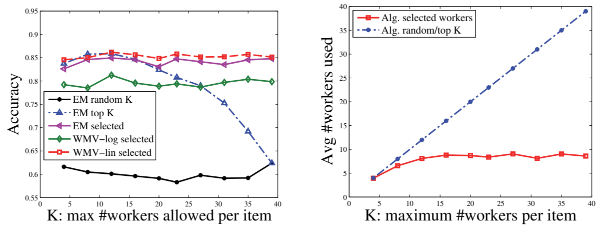
Although this combinatorial problem is neither sub-modular nor super-modular, we show it can be exactly solved with a linearithmic time algorithm shown in Algorithm 1 with global optimality guarantee (see Theorem 4 in our full paper (Li and Liu 2015)).

Algorithm 1 Worker selection algorithm

```
Input: Worker pool $\Omega = [M]$ and estimated reliabilities $\{\hat{w}_i\}_{i\in\Omega}$ from $n$ control questions; Number of label classes $L$; Constraint: no more than $K$ workers per item.
$x_i \leftarrow (L\hat{w}_i-1)^2 - L^2\hat{\mathrm{var}}(\hat{w}_i)$, $\forall i \in \Omega$ , and sort $\{x_i\}_{i\in\Omega}$ in descending order so that $x_{\sigma(1)} \geq x_{\sigma(2)} \geq \ldots \geq x_{\sigma(M)}$, where $\sigma$ is a permutation of $[M]$.
$B \leftarrow \min(K, M)$,   $g_1 \leftarrow x_{\sigma(1)}$   and   $F_1 \leftarrow g_1$.
for $k$ from 2 to $B$ do
    $g_k \leftarrow g_{k-1} + x_{\sigma(k)}$   and   $F_k \leftarrow \frac{g_k}{\sqrt{k}}$.
end for
$k^* \leftarrow \min\left\{\operatorname{argmax}_{1\leq k\leq B} F_k\right\}$.
Output: The selected subset of workers $S^\star \leftarrow \{\sigma(1), \sigma(2), \cdots, \sigma(k^*)\}$.
```

Experimental results

We test our algorithm on synthetic data and a set of real-world data sets in the full paper (Li and Liu 2015); here we only report the result on the *Bluebird dataset* collected by Welinder et al. 2010 due to space limit. In this dataset, 39 workers are asked if a presented image contains two bird spices: Indigo Bunting or Blue GroBeak. There are 108 images in total. We compared with the following algorithms: WMV-linear on the top K workers with the best reliabilities (`WMV top K`), WMV-linear on the worker set $S^\star$ selected by Algorithm 1 (`WMV-lin selected`), and WMV with log ratio weights $\log((L-1)\hat{w}_i/(1-\hat{w}_i))$ on the selected worker set $S^\star$ (`WMV-log selected`), the EM algorithm on randomly selected K workers (referred as `EM random K`), EM on the top K workers ranked (`EM top K`) and EM on the worker set $S^\star$ selected (`EM selected`).



(a) (b)

Figure 1: Performance comparison on the bluebird dataset.

The results of different algorithms are shown in Figure 1(a), where we find that `WMV-lin selected`, `WMV-log selected` and `EM selected`, all based on our worker selection algorithm, achieve better performance than EM based on the top K or the random selected workers when K is large. This shows that aggregation based on inputs from selected workers not only saves budget but also maintains good performance. Figure 1(b) shows that our algorithm selects only a small number of worker (< 10) even when the budget K is large; this indicates that a small group of well selected workers is good enough for the crowdsourcing task.

Conclusion

We studied the problem of selecting a subset of crowd workers to achieve the best accuracy for crowdsourcing labeling tasks. We demonstrated that our worker selection algorithm can simultaneously minimize the number of selected workers and minimizing the prediction error rate, achieving the best in terms of both cost and efficiency. For future directions, we are interested in developing better selection algorithms based on more advanced label aggregation algorithms such as EM, or more complex probabilistic models.

References

Dawid, A., and Skene, A. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society.* 28(1):20–28.

Karger, D.; Oh, S.; and Shah, D. 2011. Iterative learning for reliable crowdsourcing systems. In *NIPS*.

Li, H., and Liu, Q. 2015. Cheaper and better: Selecting good workers for crowdsourcing. Technical report.

Li, H.; Yu, B.; and Zhou, D. 2013. Error rate bounds in crowdsourcing models. *arXiv preprint arXiv:1307.2674*.

Liu, Q.; Ihler, A.; and Steyvers, M. 2013. Scoring workers in crowdsourcing: How many control questions are enough? In *NIPS*.

Liu, Q.; Peng, J.; and Ihler, A. 2012. Variational inference for crowdsourcing. In *NIPS*.

Mannes, A. E.; Soll, J. B.; and Larrick, R. P. 2013. The wisdom of small crowds. In *Technical Report*.

Welinder, P.; Branson, S.; Belongie, S.; and Perona, P. 2010. The multidimensional wisdom of crowds. In *NIPS*.