Optimal Worker Quality and Answer Estimates in Crowd-Powered Filtering and Rating

Akash Das Sarma
Stanford University
akashds@stanford.edu,

Aditya G. Parameswaran
University of Illinois (UIUC)
adityagp@illinois.edu

Jennifer Widom
Stanford University
widom@cs.stanford.edu

Introduction

We consider the problem of optimally filtering (or rating) a set of items based on predicates (or scoring) requiring human evaluation. Filtering and rating are ubiquitous problems across crowdsourcing applications. We consider the setting where we are given a set of items and a set of worker responses for each item: yes/no in the case of filtering and an integer value in the case of rating. We assume that items have a true inherent value that is unknown, and workers draw their responses from a common, but hidden, error distribution. Our goal is to simultaneously assign a ground truth to the item-set and estimate the worker error distribution. Previous work in this area (Raykar and Yu; Whitehill et al.) has focused on heuristics such as Expectation Maximization (EM), providing only a local optima guarantee, while we have developed a general framework that finds a maximum likelihood solution. Our approach extends to a number of variations on the filtering and rating problems.

Overview of our Approach

We are given a set of items $\mathbf{I}$ and a *worker response matrix* M, where $M(I)$ is the set of worker responses given to an item $I \in \mathbf{I}$. We assume that all workers draw their responses from a common (discrete) *error distribution*, p, where $p(i, j)$ is the probability that a worker responds i to an item with true value j. In the filtering case, $i, j \in \{0, 1\}$ while for a rating problem with R buckets, $i, j \in \{1, 2, \ldots, R\}$. Although filtering can be treated as a special case of rating, we consider it separately, as its analysis yields useful insights we build upon for the more difficult rating problem.

Our goal is to simultaneously estimate the true values of each item as well as the worker error distribution p, such that the probability of seeing the response matrix M under these assignments is maximized. Consider the filtering problem. We can compute the likelihood of any *mapping* of values to items $f : \mathbf{I} \to \{0, 1\}$ and a worker error distribution p, $P(M|f, p)$, given the response matrix M. Thus, our goal is to find $\text{argmax}_{f,p}\, P(M|f, p)$. A naive solution would be to look at every possible mapping f', compute $p' = \text{argmax}_p\, P(M|f', p)$ and $P(M|f', p')$, and choose the f' maximizing this likelihood value. The number of such mappings, $2^{|\mathbf{I}|}$, is however exponentially large. Our algorithm, based on two simple insights, greatly prunes this search space, making an exhaustive evaluation on the remaining mappings possible.

First, we are assuming (for now) that individual workers are indistinguishable, so we observe that items with the exact same set of worker responses are also indistinguishable. This allows us to bucket items based on their observed response set. For the filtering problem, suppose there are m worker responses for each item. Then we have $m + 1$ buckets, starting from m "yes" (or 1) and zero "no" (or 0) responses, down to zero yes and m no responses. We prune the set of item-value mappings by only considering those mappings that give the same value to all items in a common bucket.

Second, we observe that buckets have an inherent ordering. If workers are better than random, we intuitively expect items with more yes responses to be more likely to have actual value 1 than items with fewer yes responses. Ordering buckets by the number of yes responses, we have $(m, 0) \to (m - 1, 1) \to \ldots \to (1, m - 1) \to (0, m)$, where bucket $(m - j, j)$ contains all items that received $m - j$ yes responses and j no responses. We eliminate all mappings that give a value of 1 to a bucket on the right while assigning a value of 0 to a bucket on the left. We formalize this intuition as a *dominance relation*, or ordering on buckets, $(m, 0) > (m - 1, 1) > \ldots > (1, m - 1) > (0, m)$, and only consider mappings where dominating buckets receive a value not lower than any of their dominated buckets.

We consider the space of mappings satisfying our above bucketizing and dominance constraints, and call them *dominance-consistent mappings*. It is easy to see that in our filtering problem we have just $m + 2$ dominance-consistent mappings. We have two trivial dominance-consistent mappings, corresponding to giving every item the same value of either 1 or 0. There are exactly m other dominance-consistent mappings, each corresponding to choosing a *cut-point* $1 \leq j \leq m$, and giving all items in buckets $\{(m, 0), (m - 1, 1), \ldots, (m - j + 1, j - 1)\}$ a value of 1 and items in buckets $\{(m - j, j), (m - j - 1, j + 1), \ldots, (0, m)\}$ a value 0. We can prove that the maximum likelihood mapping from this small set of mappings is in fact a global maximum likelihood mapping across the space of all possible "reasonable" mappings: mappings corresponding to better than random worker behaviour. This result, while intuitive, involves

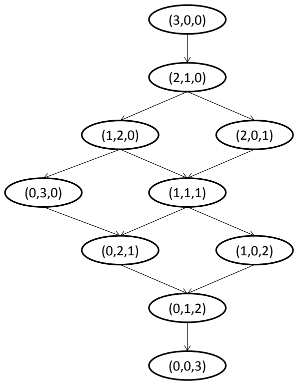

Figure 1: Dominance-DAG for 3 workers and scores in $\{1, 2, 3\}$

a fairly complex and lengthy proof. Our theorem states that if D is the space of bucketized, dominance-consistent mappings, and F is the space of all reasonable mappings, then $\max_{f \in D} P(M|f) = \max_{f \in F} P(M|f)$.

We now generalize our idea of bucketized, dominance-consistent mappings to the rating problem. The primary difference is that instead of a strictly ordered chain of buckets, we have a partial order describing the dominance constraint. We explain the generalization with the help of Figure 1. The figure shows the dominance DAG corresponding to the partial ordering among buckets when 3 independent workers respond to every item, each with a score from $\{1, 2, 3\}$. In the DAG, a node (i, j, k) represents the bucket corresponding to items where i workers give a score of 3, j workers give a score of 2 and k workers give a score of 1. Since we have 3 responses per item, $i + j + k = 3$. We naturally expect items in bucket $(3, 0, 0)$ to receive the highest rating across buckets. Similarly, we expect bucket $(1, 2, 0)$ to receive at least as high a rating as bucket $(1, 1, 1)$, although it cannot directly be compared against bucket $(2, 0, 1)$, and so on. As with the filtering problem, we can prove that an exhaustive search of the dominance-consistent mappings under this dominance DAG constraint gives us a global maximum likelihood mapping across a much larger space of reasonable mappings.

Contributions and Future Work

We have developed a simple, intuitive algorithm to solve for the maximum likelihood estimate of item-value mappings and worker error distributions. Although in the previous section we only discussed simple versions of filtering and rating, our framework generalizes to harder variants of similar problems, such as settings where different workers have different error distributions, or different items receive different number of worker responses.

While our algorithm guarantees optimal likelihood mappings, we are also interested in other metrics that measure the quality of our predicted item assignments and worker error distributions. To explore these aspects further, we ran extensive simulations with randomly-generated worker error distributions and responses, comparing the quality of our output against heuristics for different metrics. For instance, we compared the accuracy of predicted values using a distance-weighted metric, and we also measured how close our predicted worker error distribution was to the actual worker error distribution used to generate the data.

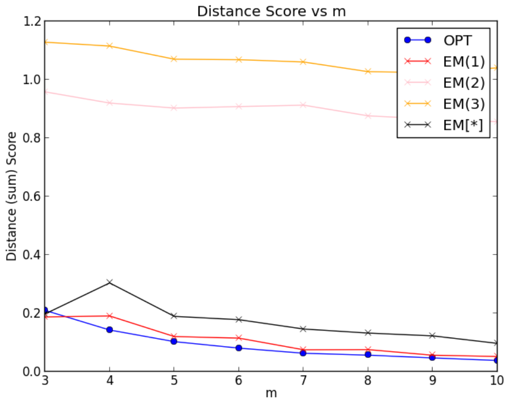

Figure 2: EMD-based score for Filtering (lower is better)

In our experiments, we plot the respective metrics (y-axis) against m, the number of worker responses per item (x-axis). Each data point is generated by averaging over 1000 different problem instances with randomly generated worker distributions and random responses to items drawn from these distributions. We show one representative plot in Figure 2, where we compare our algorithm *OPT* against the standard EM algorithm by measuring the EMD (Earth Movers Distance) between their predicted worker error distributions and the true worker distribution. We consider multiple instances of EM with a spread of different initializations $(EM(1), EM(2), EM(3))$. $EM[*]$ is a consolidated algorithm that runs each of the three instances and picks the most likely one for every given dataset. In this experiment our algorithm performs as well as the best EM approach, while also guaranteeing maximum likelihood.

As ongoing and future work, we are coping with the fact that our algorithm can be quite expensive: the number of dominance-consistent mappings can still be quite large for some problem instances. However, even in cases where our algorithm is not efficient, it provides a way to prune the search space of potential item-value mappings. A next step is to explore algorithms that use this pruned space to find maximum likelihood solutions efficiently. Also, while we have run extensive simulations, running experiments with with real human workers would likely provide additional insights.

References

Raykar, V. C., and Yu, S. 2011. Ranking annotators for crowdsourced labeling tasks. In *Advances in neural information processing systems*, 1809–1817.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J. R.; and Ruvolo, P. L. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, 2035–2043.