Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Statistical Quality Estimation for General Crowdsourcing Tasks*

Yukino Baba and **Hisashi Kashima**
The University of Tokyo

Abstract

One of the biggest challenges for requesters and platform providers of crowdsourcing is quality control, which is to expect high-quality results from crowd workers. We propose an unsupervised statistical quality estimation method for general crowdsourcing tasks with unstructured response formats such as article writing, program coding, and logo designing, which occupy the majority on most crowdsourcing marketplaces.

Introduction

One of the most challenging issues in crowdsourcing research is *quality control* to ensure the quality of crowdsourcing results, because there is no guarantee that all workers have sufficient abilities needed to complete the offered tasks at a satisfactory level of quality. A common approach to tackle this problem is to introduce redundancy, that is, to request multiple workers to work on the same tasks and aggregate their responses by applying majority voting or more sophisticated statistical aggregation techniques. The statistical quality control methods consider the characteristics of each worker or task, such as the ability of each worker and the difficulty of each task (Dawid and Skene 1979; Whitehill et al. 2009).

One serious disadvantage of these methods is that most of the existing approaches assume that the response spaces are *structured*. Binary questions (e.g., yes-or-no questions) and multiple-choice questions (e.g., five-point ratings) are typical examples where voting-like strategies work, or we can apply averaging to real-valued questions. Unfortunately, these approaches are not applicable for tasks with *unstructured response formats*, such as article writing and logo design tasks, where we cannot expect an agreement of two outputs. Most of the crowdsourcing tasks fall into this category.

One natural approach to quality estimation of artifacts for general unstructured response tasks is to employ a *two-stage workflow* as shown in Figure 1, consisting of a *creation stage* followed by a *review stage*. In the creation stage, several crowd workers (which we call *authors*) are assigned to several unstructured response tasks. Then, their artifacts proceed to the review stage, where each of them is reviewed by multiple crowd workers (called *reviewers*). The review tasks are usually casted as multiple-choice questions (such as 'Excellent,' 'Good,' 'Average,' 'Fair,' and 'Poor'). Although it is quite difficult to estimate the quality of the artifacts directly from themselves, introducing the review stage enables us to indirectly estimate the quality from the review scores, and to distinguish high-quality results from the others. For example, Zaidan and Callison-Burch (2011) applied the two-stage workflow; however, their approaches are supervised so that they require extensive domain knowledge including feature representation of artifacts and gold standard scores.

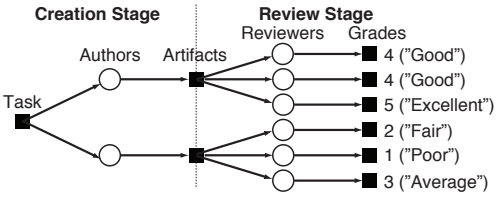

Figure 1: Example of a two-stage workflow comprising a creation stage and a review stage.

We propose an unsupervised statistical method to estimate the quality of artifacts of general unstructured response tasks using the framework of the two-stage workflow. We introduce a two-stage generative model (Figure 2). The creation stage models a generative process of the true artifact quality, where both the ability and the task-dependent performance of an author affect the quality of an artifact. The review stage models the generative process of the grade labels given by reviewers, where each reviewer first determines a latent quality score for a given artifact based on their bias and contextual preference, and then the observed grade label is generated through the graded response model (Samejima 1969) used in the item response theory.

Problem Setting

Let us assume that there is a set of general crowdsourcing tasks $\mathcal{T}$, and let $\mathcal{A}_t$ denote a set of crowd authors assigned to

*The full version of this paper appeared in the Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (Baba and Kashima 2013).

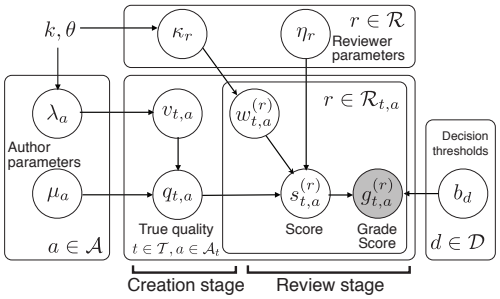

Figure 2: Graphical model of our proposed two-stage model

a task $t \in \mathcal{T}$. In the creation stage, each author $a \in \mathcal{A}_t$ creates an artifact for a task t. We denote the (unknown) quality of the artifact by $q_{t,a} \in \mathbb{R}$. In the review stage, a set of crowd reviewers $R_{t,a}$ is assigned to evaluate the quality of the artifact created by author a for task t. The evaluation by a reviewer $r \in R_{t,a}$ is given as a grade label $g_{t,a}^{(r)}$ from a set of grade labels $\mathcal{D} = \{1, 2, \cdots, n\}$. Our goal is to estimate the set of the true qualities of the artifacts $\{q_{t,a}\}_{t \in \mathcal{T}, a \in \mathcal{A}_t}$, given the set of the observed grade labels $\{g_{t,a}^{(r)}\}_{t \in \mathcal{T}, a \in \mathcal{A}_t, r \in \mathcal{R}_{t,a}}$.

Two-Stage Modeling of General Crowdsourcing Tasks

To estimate the true quality $q_{t,a}$ of the artifact created by author a for task t, we introduce a two-stage generative model, where the first stage models the generation of the artifact of quality $q_{t,a}$, and the second stage models the generation of the grade label $g_{t,a}^{(r)}$ given by reviewer r to the artifact. Figure 2 shows the graphical model of our grade label generation process. $\mu_a \in \mathbb{R}$ denotes the ability of the author $a \in \mathcal{A}$, and $1/\lambda_a \in \mathbb{R}^+$ denotes the variance of the artifact-specific noise $v_{t,a} \in \mathbb{R}$ for the pair of the task t and the author a. The true quality $q_{t,a}$ of the output is given as the sum of μ_a and $v_{t,a}$. $\eta_r \in \mathbb{R}$ denotes the evaluation bias of the reviewer $r \in \mathcal{R}$, and $1/\kappa_r \in \mathbb{R}^+$ denotes a variance of the contextual preference $w_{t,a}^{(r)} \in \mathbb{R}$ for the artifact created by the author a for the task t. The quality score $s_{t,a}^{(r)}$ is the sum of η_r, $w_{t,a}^{(r)}$, and the true quality $q_{t,a}$, which results in the observed grade $g_{t,a}^{(r)} \in \{1, 2, \ldots, n\}$ through the graded response model (Samejima 1969) with threshold parameters $\{b_d\}_d$. k and θ are hyper-parameters. We introduce prior distributions on the model parameters and apply the maximum a posteriori (MAP) inference to estimate the artifact quality as well as the other parameters.

Experiments

We conduct experiments using logo designing tasks, image description tasks, and language translation tasks on a commercial crowdsourcing platform. We compare our proposed two-stage model with two aggregation methods: majority voting and the modified Dawid-Skene model (Raykar and Yu 2011). We calculated the correlation coefficients between the estimated artifact quality scores and the ground truth grades. We also evaluated nDCG@1, which is defined as the ratio of the true quality of the estimated best artifact to that of the true best artifact. Since we could not know the "ground truths," we simulated the ground truth scores using majority voting with sufficiently many labels.

In most cases, our proposed two-stage model achieved statistically significantly higher performance over the other methods. In particular, when the number of reviewers is small, our method showed large improvements. It is notable that our model performed better even in such cases where we had only one reviewer and therefore the voting-like strategies do not work. This is because our model incorporates the creation stage with the ability parameters of authors for making the most of available information. Only in the language translation task, the simple majority voting performed the best in terms of the correlation measure. This is partly explained by the distribution of the reviewer abilities. While the reviewer abilities widely distribute in the design task and the description task, those in the translation task skew to large positive values, which implies the majority of the reviewers are reliable.

Conclusion

We proposed an unsupervised statistical method to estimate the quality of the artifacts for a general crowdsourcing tasks with unstructured response formats. Experimental results showed the advantage of our two-stage model compared with some existing label aggregation methods, especially when limited numbers of reviewers and authors are available, which implies that the proposed method can deliver high-quality crowdsourcing results with lower costs.

Acknowledgments

This research was supported by the Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program).

References

Baba, Y., and Kashima, H. 2013. Statistical quality estimation for general crowdsourcing tasks. In *Proc. of KDD*.

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

Raykar, V. C., and Yu, S. 2011. Ranking annotators for crowdsourced labeling tasks. In *Proc. of NIPS*.

Samejima, F. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*.

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proc. of NIPS*.

Zaidan, O. F., and Callison-Burch, C. 2011. Crowdsourcing translation: professional quality from non-professionals. In *Proc. of ACL-HLT*.