On Reliable Crowdsourcing and the Use of Ground Truth Information

Qiang Liu
Dept. of Computer Science
UC Irvine

Mark Steyvers
Dept. of Cognitive Sciences
UC Irvine

John W. Fisher, III
CSAIL
MIT

Alexander Ihler
Dept. of Computer Science
UC Irvine

Abstract

Using some expert labels or control questions with known answers may significantly improve the reliabilities of crowdsourcing systems, but with high cost that may not be feasible at large scale. This raises important problems on how to optimally integrate such additional, expensive but accurate information with the cheap but unreliable crowd labels, addressing which is critical to build more reliable and scalable crowdsourcing systems in the future. We give an overview of some related open problems, along with some current work.

The crowdsourcing paradigm has emerged as a powerful approach for collecting human knowledge and intelligence at large scale. One of its underlying philosophies is the *wisdom of crowds* phenomenon (Surowiecki, 2005): properly combined, a group of untrained people can significantly outperform average individuals, and sometimes even rival domain experts. Unfortunately, the (often anonymous) crowd workers have unknown and highly diverse levels of expertise and possible biases, making it a critical problem to monitor workers' performance and optimally combine the results of the crowd.

One common *"supervised"* solution is to take advantage of additional ground truth information or input from domain experts. This can be done by evaluating workers' performance using a number of *control items*, which have been pre-labeled with known answers and "seeded" into workers' task sets without telling them. CrowdFlower, for example, provides interfaces and tools to allow requesters to explicitly specify and analyze a set of control items (someteimes called *gold data*). Alternatively, a set of items can be selected to be labeled by domain experts (whom we assume give correct answers) *after* the crowdsourcing phase is complete; this enables better, more adaptive decisions about how many and which items should be checked by experts, allocating the experts' work budget more effectively when expert labels are expensive. Furthermore, one may create more flexible procedures that iteratively acquire labels from either experts or the crowd in an adaptive manner until some stopping criterion (such as estimated quality) is satisfied.

Perhaps surprisingly, some *"unsupervised"* solutions are able to evaluate workers' performance without any ground truth information (e.g., Dawid and Skene, 1979, Liu, Peng, and Ihler, 2012, Zhou et al., 2012). These methods often work by building and conducting statistical inference over a joint probablistic model of the workers and labels, and rely on the idea of scoring workers by their agreement with other workers, assuming that the majority of workers (overall) are correct. Some work indicates that the worker reliabilities estimated by these unsupervised methods can be almost as good as those estimated when the true labels of all items are known (e.g., Lee et al., 2012), sometimes with theoretical guarantees (e.g., Karger, Oh, and Shah, 2011).

The success of these unsupervised methods seemingly eliminates the need to have ground truth information, at least in some cases, and raises the question of when and how much such ground truth can help. A deep understanding of these issues will become critical in the future, since the cost of ground truth may become a bottleneck for scalability.

Optimal Usage of Control Items

Some aspects of these problems have been studied recently by Liu, Steyvers, and Ihler (2013), which focuses on using control items to evaluate workers' performance and improve the resulting label aggregation algorithms. They study the problem of how many control items should be used, which demonstrates a clear trade-off: having workers answer more control items gives better estimates of their performance, but leaves less resources for the target items whose answers are of real interest; on the other hand, using too few control items gives poor estimates of workers' performance, also leading to bad results. Liu, Steyvers, and Ihler (2013) shows that the optimal number of control items depends on how the control items are used to help the prediction. They study two basic types of aggregation methods that incorporate the control items in different ways:

- *Two-stage methods*, which first score the workers based on their performance on control items, then use these scores to improve aggregation on the remaining target items (whose answers are unknown and of direct interest), by eliminating or down-weighting workers that perform badly on control items.
- *Joint inference methods,* which simultaneously estimate workers' reliabilities and the answers on target items us-

ing methods such as maximum likelihood or Bayesian inference on joint probability models over the crowd answers and the true labels.

Liu, Steyvers, and Ihler (2013) provides simple scaling rules for the optimal number of control items assigned to each worker: either linearly with or on the order of the square root of the total number of items per worker, depending on the type of the aggregation method. They also show that the joint inference method is very sensitive to model misspecificaiton: although it requires very few control items and provides high accuracy if the model assumptions hold, it tends to degenerate significantly or requires many more control items if the model and data are mismatched.

However, many other important problems are still open. In practice, the control items are often used (sometimes implicitly) as validation sets to support higher level decisions, like model selection, hyper-parameter tuning, and even algorithm and workflow design. These uses can have significant impact on performance, and raises many critical questions, such as: What is the optimal number of control items for different levels of decisions? How to allocate items across different levels? How much can performance be boosted? A principled study would greatly enhance our abilities to build better crowdsourcing systems in the future.

Integrating the Crowd with Experts

Since labels from domain experts are much more expensive then those from the crowd, it is important to optimally allocate these expert resources. To this end, it is more flexible to acquire expert labels *after* the crowd labels are collected, making it possible to leverage the information from the crowd and make better decisions on how many and which items should be checked by experts. Furthermore, one could design even more flexible procedures that adaptively acquire labels from experts and the crowd until either a budget limit or accuracy criterion are satisfied. This would enable platforms that efficiently integrate the crowd and domain experts automatically and at large scale, but involves solving a complex, online decision process over (1) whether to add expert labels or some amount of crowd labels, (2) which items to label, and (3) when to stop the process. Although there exist a large body of work on online decision making in crowdsourcing (e.g., Slivkins and Vaughan, 2013), the particular setting of allocating expert effort appears under-studied.

Phase Transition. In some preliminary work, we find that the optimal allocation of expert labels undergoes a transition between two phases:

- *Global Phase.* When the existing number of expert labels is small, the characteristics of the workers are largely undetermined. In this case, it is optimal to acquire expert labels on the *"most influential"* items whose true answer will most improve the evaluation of workers' characteristics. These influential items are usually those labeled by many workers, e.g., hub nodes in the graph encoding assignment between the items and workers. Since improving worker evaluation then helps in evaluating their assigned items, the overall improvement introduced by the expert labels is very significant (a "snowball" effect).
- *Myopic Phase.* When the number of expert labels is sufficiently large, the characteristics of the workers are well estimated, and the snowball effect tends to saturate. In this stage, it appears optimal to acquire expert labels on the *most uncertain* items. In this case, the error rate decreases only linearly with the number of expert labels, a much lower return on investment than in the global phase.

The (usually highly expensive) expert labels in the myopic phase may not be worth their cost, since they essentially affect only single items and so should be replaced by cheaper crowd labels. We argue that one should stop acquiring expert labels at the myopic phase and switch to acquiring more inexpensive labels from new crowd workers, returning the system back to the global phase due to the increase in workers' uncertainty. This provides some initial intuition on how to construct systems that automatically switch between crowd and expert labels and trade off reliability and cost optimally.

Conclusions

We discuss the use and need for ground truth labels in crowdsourcing systems, the minimization of which is critical to the scalability and usefulness of crowdsourcing in the future. By understanding the optimal balance between ground truth and crowd labels, and the regimes in which acquiring expert labels are most useful, we can begin to design systems that automate the balance between these two sources and even adaptively switch between them to minimize cost while maximizing quality and reliability. Solving these open problems is thus a key component of using crowdsourcing effectively at scale.

References

Dawid, A., and Skene, A. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics* 20–28.

Karger, D.; Oh, S.; and Shah, D. 2011. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems (NIPS)*, 1953–1961.

Lee, M. D.; Steyvers, M.; de Young, M.; and Miller, B. 2012. Inferring expertise in knowledge and prediction ranking tasks. *Topics in cognitive science* 4(1):151–163.

Liu, Q.; Peng, J.; and Ihler, A. 2012. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems (NIPS)*, 701–709.

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

Slivkins, A., and Vaughan, J. W. 2013. Online decision making in crowdsourcing markets: Theoretical challenges (position paper). *arXiv preprint arXiv:1308.1746*.

Surowiecki, J. 2005. *The wisdom of crowds*. Anchor.

Zhou, D.; Platt, J.; Basu, S.; and Mao, Y. 2012. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems (NIPS)*, 2204–2212.