Efficiency of active learning for the allocation of workers on crowdsourcing classification tasks[*]

Edoardo Manino[†] Long Tran-Thanh[†] Nicholas R. Jennings[‡]

Abstract

Crowdsourcing has been successfully employed in the past as an effective and cheap way to execute classification tasks and has therefore attracted the attention of the research community. However, we still lack a theoretical understanding of how to collect the labels from the crowd in an optimal way. In this paper we focus on the problem of worker allocation and analyse two active learning policies proposed in the empirical literature. In the end we are able to show their equivalence and prove their superiority over an existing theoretical result by Karger et al. [2011].

Introduction and background

Crowdsourcing has emerged as an effective way of completing a large number of simple and repetitive tasks by hiring workers from an online community in exchange for a small payment. Despite the unreliability of the single members, as a whole these crowds have been proved capable of producing valuable solutions at a fraction of the cost of traditional methods. A common way of interacting with the workers is to present them with a series of classification tasks and use their answers to infer the correct labels. This model has been successfully employed for various purposes including image labeling [Sorokin and Forsyth, 2008], ballot voting in complex workflows [Dai et al., 2011] and filtering [Parameswaran et al., 2012]. Even so, we still lack a principled approach towards the management of the crowd during the label collection process, and how to efficiently allocate the workers on the available tasks.

One of the main obstacles to the realisation of such a goal is the heterogeneity in the quality of the answers collected from the crowd [Ipeirotis et al., 2010]. As a large number of them may be incorrect, the employer is forced to require multiple labels for each task in order to come up with a reliable conclusion. Over the past decades several statistical methods have been proposed to efficiently aggregate those labels and reduce the number of misclassified items [Dawid and Skene, 1979, Liu et al., 2012]. However, these methods are designed to be used after the labels have been collected from the crowd, and give no indication on how to optimise the collection process itself.

This second question has recently received more attention in the literature [Slivkins and Vaughan, 2013]. On the theoretical side, Karger et al. [2011] analyse a scenario with homogeneous tasks and propose to assign the same number of workers on each task, showing that this approach is order-optimal to the best allocation knowing the quality of the workers in advance. Conversely, when the tasks are heterogeneous, Ho et al. [2013] advocate the use of online primal-dual techniques to improve the efficiency of the allocation. However, the authors of these two papers fail to use the labels provided by the crowd as a potential source of information to improve the allocation process.

In contrast, such piece of information has been extensively used in more empirical-oriented works on crowdsourcing [Kamar et al., 2012, Chen et al., 2013]. In particular Welinder and Perona [2010] propose an algorithm that computes the confidence in the classifications after every new label, and allocates more workers on the most uncertain tasks. Simpson and Roberts [2014] address the same problem from a different perspective and introduce a greedy algorithm to maximise the amount of information

[*] This research is funded by the UK Research Council project ORCHID, grant EP/I011587/1

[†] University of Southampton

[‡] Imperial College, London

collected from the crowd. Notably these two approaches belong to the family of *active learning* policies [Settles, 2010], but the authors provide no theoretical guarantee on their performances.

In this paper we fill this gap by analysing these two empirical approaches from the theoretical perspective. Specifically we show that for binary classification tasks the two active learning policies are equivalent, and we prove that, in an ideal scenario, they achieve a greater efficiency than the uniform allocation proposed by Karger et al. [2011].

An active learning policy for worker allocation

We analyse here a setup where we have to discover the true binary label $\ell_i^* \in \{+1, -1\}$ of a whole set of independent, homogeneous tasks $i \in [1, M]$ given a total budget B. The interaction with the crowd proceeds in a series of rounds, where a new worker j becomes available, is allocated to a task i under a fixed payment $c = 1$, and produces a label ℓ_{ij} which is correct with probability p_j. Additionally, we assume the presence of an oracle that reveals the value of p_j as soon as the round begins. We predict the label of each task as $\hat{\ell}_i = \operatorname{sign} \Phi_i$ where $\Phi_i = \sum_j \ell_{ij} w_j$ is the weighted sum of the labels collected for i so far, and $w_j = \log(p_j/(1-p_j))$ are the optimal weights [Nitzan and Paroush, 1982].

We can now adapt the worker allocation policy of Simpson and Roberts [2014] to our setup by choosing the task i that maximises the amount of information collected at each round. To do so we interpret the effect of the next incoming label as a random variable $x_j = \pm w_j$ which will modify the current posterior probability $\mathbb{P}(\ell_i^* = +1) = \exp(\Phi_i)/(\exp(\Phi_i)+1)$, and measure its information gain as the Kullback-Leibler divergence between the two distributions:

$$\mathcal{I}(i, x_j) = \frac{x_j \exp(\Phi_i + x_j)}{\exp(\Phi_i + x_j) + 1} + \log\left(\frac{\exp(\Phi_i) + 1}{\exp(\Phi_i + x_j) + 1}\right) \tag{1}$$

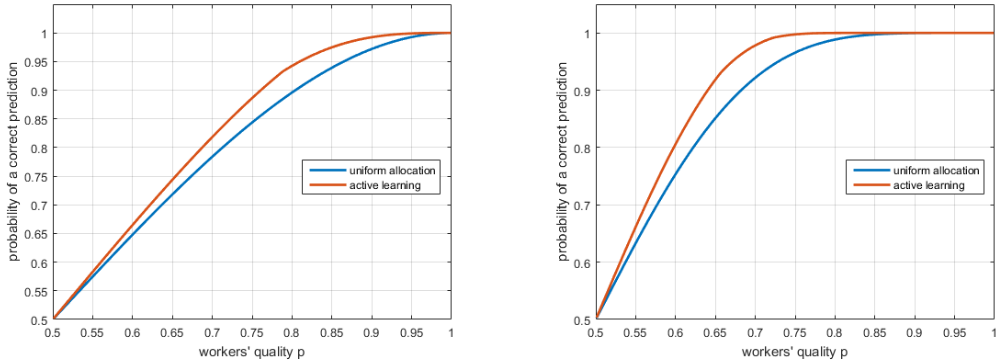
As the sign of x_j is unknown a priori, we need to choose the task i that maximises Equation 1 in expectation, i.e. $i^* = \operatorname{argmax}_i(\mathbb{E}_{x_j}(\mathcal{I}(i, x_j)))$. Fortunately it is possible to perform such computation explicitly by taking into account the current posterior on i and the probability p_j of observing the correct label. Moreover it can be shown that the expected information gain is a symmetric function around its maximum in $\Phi_i = 0$ for any $w_j \neq 0$ (we omit the derivation here for lack of space). As a consequence we can make the optimal greedy decision using the alternative formula $i^* = \operatorname{argmin}_i(|\Phi_i|)$ which corresponds to selecting the task with the most uncertain classification, a popular active learning policy [Lewis and Gale, 1994, Welinder and Perona, 2010].

Comparison with Karger's uniform allocation

In order to better understand the properties of the policy introduced above, let us restrict our attention to the case where all the workers have the same probability $p_j = p$ and define $q = 1 - p$. Furthermore let us denote $\Phi_B = \min_i(|\Phi_i|) = |k \log(p/q)|$ as the minimum confidence threshold at the end of the crowdsourcing process, for some $k \in \mathbb{N}$.

From the point of view of a single task i, the active learning policy keeps allocating new workers on i in short bursts of activity, until $|\Phi_i|$ reaches at least Φ_B. We can interpret this behaviour as a bounded random walk in a Markov domain with states in the ordered set $S = (-k, \ldots, 0, \ldots, +k)$, where the first and last state have the special property of terminating the random walk. As far as the transition probabilities are concerned, we can move only between neighbouring states with probability p of going towards the correct label (which we conventionally set to $\ell_i^* = +1$) and q of going in the opposite direction. We can summarise them in a sparse tridiagonal matrix T_k which has the first and last row empty, as in the example that follows:

$$T_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ 0 & q & 0 & p & 0 \\ 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{2}$$



(a)

(b)

Figure 1: comparison between the uniform allocation and the active learning policies under the same budget constraints, (a) $R_{uni} = 3$ votes per task and (b) $R_{uni} = 11$.

With a number of tasks $M \to \infty$, the average number of labels R_{act} consumed by the random walk to reach confidence Φ_B converges to its expected value:

$$\mathbb{E}(R_{act}) = \sum_{r=1}^{\infty} r \left(e_{\emptyset}^{T} (T_k)^r (e_{-k} + e_{+k}) \right) = \frac{(2P_{act} - 1)\Phi_B}{(2p-1)\log(p/q)} \tag{3}$$

where e_s is an indicator vector for the state $s \in S$ and $P_{act} = \exp(\Phi_B)/(\exp(\Phi_B)+1)$ is the probability of making a correct prediction.

Finally, we can compare the efficiency of this active learning policy with Karger's uniform allocation. Let us assume that our budget B is such that we can uniformly allocate an odd number of workers $R_{uni} \geq 3$ on each task: the probability of having a majority on the correct class is thus $P_{uni} = \sum_{r=\lceil R_{uni}/2 \rceil}^{R_{uni}} p^r q^{R_{uni}-r}$. Under the same budget constraints we can use Equation 3 to compute the probability P_{act} of predicting the correct class for the active learning policy. The results in Figure 1 show that the value of P_{act} is consistently above P_{uni} for any $p \in (0.5, 1)$, hence proving that this policy is more efficient than the uniform one in allocating a given budget B.

Conclusions

In this paper we analysed an active learning policy for the allocation of workers on binary classification tasks, and proved its equivalence to two approaches from the empirical literature on crowdsourcing. Furthermore we showed that, when the quality of the crowd is homogeneous, this policy outperforms Karger's uniform allocation in terms of prediction accuracy. Future works include extending the proof to crowds with heterogeneous workers, and addressing a more realistic scenario where we can only access noisy estimates of their quality.

References

Xi Chen, Qihang Lin, and Dengyong Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 64–72, 2013.

Peng Dai, Mausam, and Daniel S. Weld. Artificial Intelligence for Artificial Artificial Intelligence. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pages 1153–1159, 2011.

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

Chien-ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. Adaptive Task Assignment for Crowdsourced Classification. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 534–542, 2013.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, page 64, 2010.

Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, volume 1, pages 467–474, 2012.

David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-Optimal Crowdsourcing using Low-rank Matrix Approximations. In *Proceedings of the 49th Annual Conference on Communication, Control, and Computing (Allerton)*, pages 284–291, 2011.

David D Lewis and William A Gale. A Sequential Algorithm for Training Text Classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

Qiang Liu, Jian Peng, and Alexander T Ihler. Variational Inference for Crowdsourcing. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 692–700, 2012.

Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–297, 1982.

Aditya G. Parameswaran, Hector Garcia-Molina, Hyunjung Park, Neoklis Polyzotis, Aditya Ramesh, and Jennifer Widom. CrowdScreen: Algorithms for Filtering Data with Humans. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 361–372, 2012.

Burr Settles. Active Learning Literature Survey. *Machine Learning*, 15(2):201–221, 2010.

Edwin Simpson and Stephen Roberts. Bayesian Methods for Intelligent Task Assignment in Crowdsourcing Systems. In *Scalable Decision Making: Uncertainty, Imperfection, Deliberation.* 2014.

Aleksandrs Slivkins and Jennifer Wortman Vaughan. Online Decision Making in Crowdsourcing Markets: Theoretical Challenges. *ACM SIGecom Exchanges,*, 12(2):4–23, 2013.

Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. *Urbana*, 51(61):820, 2008.

Peter Welinder and Pietro Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 25–32, 2010.