# Quality Control of Crowd Labeling through Expert Evaluation

**Faiza Khan Khattak**
Department of Computer Science
Columbia University
New York, NY 10027
fk2224@columbia.edu

**Ansaf Salleb-Aouissi**
Center for Computational Learning Systems
Columbia University
New York, NY 10027
ansaf@ccls.columbia.edu

## Abstract

We propose a general scheme for quality-controlled labeling of large-scale data using multiple labels from the crowd and a "few" ground truth labels from an expert of the field. Expert-labeled instances are used to assign weights to the expertise of each crowd labeler and to the difficulty of each instance. Ground truth labels for *all* instances are then approximated through those weights along with the crowd labels. We argue that injecting a *little* expertise in the labeling process, will significantly improve the accuracy of the labeling task. Indeed, empirical evaluation demonstrates that our methodology is efficient and effective as it gives better quality labels than majority voting and other state-of-art methods.

## 1 Introduction

Recently, there has been an increasing interest in *Learning from Crowd* [8], where people's intelligence and expertise can be gathered, combined and leveraged in order to solve different kinds of problems. Specifically, crowd-labeling emerged from the need to label large-scale and complex data, often a tedious, expensive and time-consuming task. While reaching for human cheap-and-fast labeling (e.g. using Mechanical Turk) is becoming widely used by researchers and practitioners, the quality and integration of different labels remain an open problem. This is particularly true when the labelers participating to the task are of unknown or unreliable expertise and generally motivated by the financial reward of the task [3].

We propose a framework called Expert Label Injected Crowd Estimation (ELICE) for accurate labeling of data by mixing multiple labels from the crowd and a few labels form the experts of the field. We assume that experts always provide ground truth labels without making any mistakes. The level of expertise of the crowd will be determined by comparing their labels to the expert labels. Similarly expert labels will also be used to judge the level of difficulty of the instances labeled by the experts. Hence, expert-labeled instances are used to assign weights to the expertise of each crowd labeler and to the difficulty of each instance. True labels for *all* instances are then approximated through those weights along with the crowd labels. We provide an efficient and effective methodology that demonstrates that injecting a *little* expertise in the labeling process, does significantly improve the accuracy of the labeling task.

This paper is organized as follows: related work is described in Section 2. In Section 3, we present our Expert Label Injected Crowd Estimation (ELICE) framework along with its clustering-based variant. Experiments demonstrating the efficiency and accuracy of our methodology are provided in Section 4. We finally conclude with a summary and future work in Section 5.

## 2 Related Work

Quite a lot of recent work has addressed the topic of *learning from crowd* [8]. Whitehill et al. [13] propose a probabilistic model of the labeling process which they call GLAD (Generative model of Labels, Abilities, and Difficulties). Expectation-maximization (EM) is used to obtain maximum likelihood estimates of the unobserved variables and is shown to outperform "majority voting." Whitehill et al. [13] also propose a variation of GLAD that uses clamping in which it is assumed that the true labels for some of the instances are known somehow which are "clamped" into the EM algorithm by choosing the prior probability of the true labels very high for one class and very low for the other. A probabilistic framework is also proposed by Yan et al. [14] as an approach to model annotator expertise and build classification models in a multiple label setting. Raykar et al. [9] proposed a Bayesian framework to estimate the ground truth and learn a classifier. The main novelty of this work is the extension of the approach from binary to categorical, ordinal and continuous labels. Another related line of research is to show that using multiple, noisy labelers is as good a using fewer expert labelers. Work in that context includes Sheng et al. [10], Snow et al. [11] and Sorokin and Forsyth[12]. In [1], the authors propose a method to increase the accuracy of the crowd labeling using the concept of active learning by choosing the most informative labels. This is done by constructing a confidence interval called "Interval Estimate Threshold" for the reliability of labeler. More recently, [15] have also developed a probabilistic method based on the idea of active learning to use the best possible labels from the crowd. In [2] active learning approach is adopted for labeling based on features instead of the instances and it is shown that this approach outperforms the corresponding passive learning approach based on features.

In our framework, expert-labeled instances are used to evaluate the crowd expertise and the instance difficulty, to improve the labeling accuracy.

## 3 Our Framework

### 3.1 Expert Label Injected Crowd Estimation (ELICE)

Consider a dataset of $N$ instances, which is to be labeled as positive (label= +1) or negative (label=-1). A subset of $n$ instances is labeled by an expert of the domain. There are $M$ crowd labelers who label all $N$ instances. Label given to the $i^{th}$ instance by the $j^{th}$ labeler is denoted by $l_{ij}$. The expertise level of a labeler $j$ is denoted by $\alpha_j$, which can have value between -1 and 1, where 1 is the score for a labeler who labels all instances correctly and -1 is the score for the labeler who labels everything incorrectly. This is because the expertise of a crowd labeler is penalized by subtracting 1 when he makes a mistake but it is incremented by 1 when he labels correctly. At the end the sum is divided by $n$. Similarly, $\beta_i$ denotes the level of the difficulty of the instance $i$ which is calculated by adding 1 when a crowd labeler labels that particular instance correctly. Sum is normalized by dividing by $M$. It can have value between 0 and 1, where 0 is for the difficult instance and 1 is for the easy one. Formulas for calculating $\alpha_j$'s and $\beta_i$'s are as follows,

$$\alpha_j = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}(L_i = l_{ij}) - \mathbf{1}(L_i \neq l_{ij})] \qquad \beta_i = \frac{1}{M}\sum_{j=1}^{M}[\mathbf{1}(L_i = l_{ij})] \tag{1}$$

where $j = 1, \ldots, M$ and $i = 1, \ldots, n$.

We infer the rest of the $(N - n)$ number of $\beta$'s based on $\alpha$'s. As the true labels for the rest of instances are not available, we try to find an approximation which we name as *expected label* (EL)

$$s_i = \frac{1}{M}\sum_{j=1}^{M}\alpha_j * l_{ij}, \qquad EL_i = \begin{cases} 1 & \text{if } s_i \geq 0 \\ -1 & \text{if } s_i < 0 \end{cases} \tag{2}$$

These expected labels are used to approximate $\beta$'s ,

$$\beta_i = \frac{1}{M}\sum_{j=1}^{M}[\mathbf{1}(EL_i = l_{ij})] \tag{3}$$

Logistic function is used to calculate the score associated with the correctness of a label, based on the level of expertise of the crowd labeler and the difficulty of the instance. This score gives us the approximation of the true labels called *inferred labels* (IL) using the following formulas

$$P_i = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{(1+exp(-\alpha_j\beta_i))} * l_{ij}, \qquad IL_i = \begin{cases} 1 & \text{if } P_i \geq 0 \\ -1 & \text{if } P_i < 0 \end{cases} \tag{4}$$

| | Dataset | Mushroom | Chess | Tic-Tac-Toe | Breast Cancer | IRIS |
|---|---|---|---|---|---|---|
| Bad Labelers | Instances | 8124 | 3196 | 958 | 569 | 100 |
| | Expert labels | 20 | 8 | 8 | 8 | 4 |
| Less than 30% | Majority Voting | 0.9922 | 0.9882 | 0.9987 | 0.9978 | 1.000 |
| | GLAD | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | GLAD with clamping | 1.000 | 1.000 | 1.0000 | 1.000 | 1.000 |
| | ELICE | 0.9998 | 1.000 | 1.000 | 0.9996 | 1.0000 |
| | ELICE with clustering | 0.9997 | 1.000 | 1.000 | 1.000 | 1.000 |
| 30 to 70% | Majority Voting | 0.3627 | 0.3702 | 0.3732 | 0.4332 | 0.3675 |
| | GLAD | 0.5002 | 0.5001 | 0.2510 | 0.5011 | 0.2594 |
| | GLAD with clamping | 0.5002 | 0.5001 | 0.2510 | 0.5011 | 0.5031 |
| | ELICE | 0.9917 | 0.9932 | 0.9961 | 0.9966 | **0.9931** |
| | ELICE with clustering | **0.9947** | **0.9944** | **0.9968** | **0.9979** | **0.9931** |
| More than 70% | Majority Voting | 0.0002 | 0.0014 | 0.0010 | 0.0023 | 0.0100 |
| | GLAD | 0.0002 | 0.0039 | 0.0013 | 0.0022 | 0.0125 |
| | GLAD with clamping | 0.0002 | 0.0039 | 0.0013 | 0.0022 | 0.0125 |
| | ELICE | 0.5643 | 0.5884 | 0.5242 | 0.5599 | 0.5451 |
| | ELICE with clustering | **0.7001** | **0.6481** | **0.6287** | **0.6249** | **0.5777** |

Table 1: Accuracy of Majority voting, GLAD (with and without clamping) and ELICE (with and without clustering) for different datasets and different rates of bad labelers.

### 3.2 Expert Label Injected Crowd Estimation with clustering

ELICE with clustering is a variation of the above method with the only difference that instead of picking the instances randomly from the whole dataset for getting expert labels, clusters are formed and then equal number of instances are chosen from each cluster and given to expert to label.

## 4 Experiments

We conducted experiments on several datasets from the UCI repository including Mushroom, Chess, Tic-Tac-toe, Breast cancer and IRIS (with this latter restricted to 2 classes). Table 1 shows a comparison of the accuracy of different methods including ours across different datasets for different rates of bad labelers. Bad labelers are assumed to make more than 80% mistakes while the rest of the labelers are assumed to be good and make at most 20% mistakes. It can be observed from the table that our methods outperform majority voting, GLAD and GLAD with clamping even when the percentage of bad labelers is increased to more than 70%. The corresponding graph for Mushroom dataset is shown in Figure1 (left). Our method also has the advantage of having a runtime, which is significantly less than that of GLAD and GLAD with clamping as shown in Figure 1 (right). In all these experiments the number of crowd labelers is 20. We have also shown that increasing the number of expert labels increases the accuracy of the result up to a certain level, as shown in 2 for the Mushroom dataset. In this case we have assumed that the bad labelers are adversarial and are making 90% mistakes.

## 5 Discussion & Future Work

We propose a new method, named ELICE, demonstrating that injecting a "few" expert labels in a multiple crowd labeling setting improves the estimation of the reliability of the crowd labelers as well as the estimation of the actual labels. ELICE is efficient and effective in achieving this task for

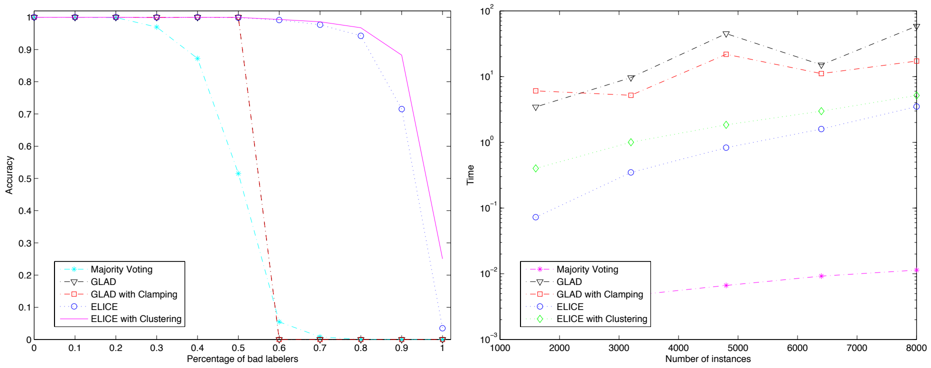

Figure 1: (Left) Number of bad labelers vs. Accuracy (Right) Number of instances vs. time with 20 expert labels for ELICE and ELICE with clustering

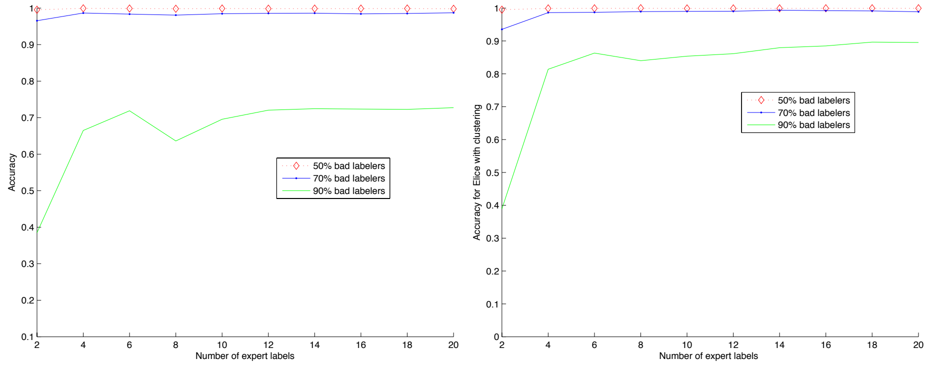

Figure 2: (Left) Number of expert labelers vs. Accuracy for ELICE (Right) Number of expert labelers vs. Accuracy for ELICE with clustering

*any* setting where crowd labeler's quality is heterogenous, that is a mix of good and bad labelers. Bad labelers can be either adversarial in the extreme case or labelers performing a large amount of mislabeling because of their lack of expertise or the difficulty of the task. Furthermore, it turned out that our estimation methodology remains reliable even in the presence of a large proportion of such bad labelers in the crowd. We show through extensive experiments that our method is robust even when other state-of-the-art methods fail to estimate the true labels. Identifying adversarial labelers is not new, and is tackled through an a priori identification of those labelers before the labeling task starts (e.g. [7]). ELICE has the ability of handling not only adversarial but bad labelers in general in an integrated way instead of identifying them separately.So far, we have assumed that crowd labelers are assigned all instances to label, which may be impractical and a waste of resources. Instead, one can think of dividing the instances into equal-sized subsets and assign a fixed number of subsets to each crowd labeler. An equal number of instances can then be drawn from each subset to be labeled by the expert. Recently, an elaborated approach of instance assignment was proposed in [5, 4, 6]. A bipartite graph is used to model the instances and labelers. The edges between them are used to describe which labeler will label which instance. The graph is assumed to be regular but not complete so as not every instance gets labeled by every labeler.

Another extension of our method is to acquire expert labels for carefully chosen instances. We propose a variant of ELICE using clustering as one way to achieve this. By clustering instances, and having the expert label an equal number of instances from each cluster, we hope to have labels for instances representatives of the whole set. Another possibility is to ask the expert to identify and label difficult instances on which crowd will likely make most mistakes. Finally, our next goal is to study theoretically the balance between the number of expert-labeled instances to acquire and crowd labelers in order to find a good compromise between good accuracy and minimum cost.

## Acknowledgments

This project is partially funded by a Research Initiatives in Science and Engineering (RISE) grant from the Columbia University Executive Vice President for Research.

## References

[1] Pinar Donmez, Jaime G. Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *KDD*, pages 259–268, 2009.

[2] G. Druck, B. Settles, and A. McCallum. Active learning by labeling features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 81–90. ACL Press, 2009.

[3] Karën Fort, Gilles Adda, and K. Bretonnel Cohen. Amazon mechanical turk: Gold mine or coal mine? *Comput. Linguist.*, 37:413–420, June 2011.

[4] David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. *Proc. of the Allerton Conf. on Commun., Control and Computing, Monticello, IL*, 2011.

[5] David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *CoRR*, abs/1110.3564, 2011.

[6] David R. Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. *Neural Information Processing Systems (NIPS), Granada, Spain*, 2011.

[7] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, Vol. 5, No. 5:411–419, 2010.

[8] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, April 2010.

[9] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 889–896, New York, NY, USA, 2009. ACM.

[10] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622, New York, NY, USA, 2008. ACM.

[11] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[12] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. *Computer Vision and Pattern Recognition Workshops*, Jan 2008.

[13] Jacob Whitehill, Paul Ruvolo, Ting fan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2035–2043. 2009.

[14] Yan Yan, Rosales Rómer, Fung Glenn, Schmidt Mark, Hermosillo Gerardo, Bogoni Luca, Moy Linda, and Dy Jennifer G. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the12th International Conference on Artificial Intelligence and Statistics*, 2010.

[15] Yan Yan, Romer Rosales, Glenn Fung, and Jennifer Dy. Active learning from crowds. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1161–1168, New York, NY, USA, June 2011. ACM.