Toward a Robust and Universal Crowd-Labeling Framework

Faiza Khan Khattak
Columbia University, New York
fk2224@columbia.edu

Abstract

One of the main challenges in crowd-labeling is to control for or determine in advance the proportion of low-quality/malicious labelers. We propose methods that estimate the labeler and data instance related parameters using frequentist and Bayesian approaches. All these approaches are based on expert-labeled instance (ground truth) for a small percentage of data to learn the parameters. We also derive a lower bound on the number of expert-labeled instances needed to get better quality labels.

1 Introduction

Crowd-labeling is the process of having a human crowd label a large dataset. It is well-known that the precision and accuracy of labeling can vary due to differing skill sets. The labelers can be *good/experienced*, *random/careless* or even *malicious*. If the proportion of malicious labelers grows too high, there is often a *phase transition* leading to a steep, non-linear drop in labeling accuracy as noted by [Karger *et al.*, 2014]. We propose methods for a robust and accurate crowd-labeling system that delays the phase transition. Our hypothesis is that using some expert-labeled instances (ground truth) can help us get insight about the labeler-ability as well as instance-difficulty, which can help in improving the accuracy of the aggregated final label. We propose a frequentist and a Bayesian approaches to learn these parameters.

2 Frequentist Approach

Based on our hypothesis, we propose the first version of Expert Label Injected Crowd Estimation (ELICE) [Khattak and Salleb-Aouissi, 2011]. We estimate the labeler-ability α and instance-difficulty β based on a few (usually 0.1% -10% of the whole dataset) expert-labeled instances as:
$\alpha_j = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}(L_i = l_{ij}) - \mathbf{1}(L_i \neq l_{ij})]$,
$\beta_i = \frac{1}{M}\sum_{j=1}^{M}[\mathbf{1}(L_i = l_{ij})]$, where L_i is the true label for instance i, l_{ij} is the label given by labeler j to instance, i, $j = 1, \ldots, M$ and $i = 1, \ldots, n$. Next β's for the rest of the data (with no expert-labels) are estimated based on α's as:

$$EL_i = sign(\frac{1}{M}\sum_{j=1}^{M}\alpha_j * l_{ij}), \beta_i = \frac{1}{M}\sum_{j=1}^{M}[\mathbf{1}(EL_i = l_{ij})] \quad (1)$$

Label aggregation is done by using logistic function (σ).
$IL_i = sign(\frac{1}{M}\sum_{j=1}^{M}\sigma(\alpha_j\beta_i) * l_{ij})$
ELICE 1 [Khattak and Salleb-Aouissi, 2011] is efficient as well as effective. It assigns high weights to the good labelers' annotation to identify the correct final labels. To further squeeze information, even from the malicious labelers, we proposed ELICE 2 [Khattak and Salleb-Aouissi, 2013]. In this method, we introduce entropy as a way to estimate the uncertainty of labeling. This provides an advantage of differentiating between good, random and malicious labelers. The aggregation method for ELICE version 2 flips the label (for binary classification case) provided by the malicious labeler thus utilizing the information that is generally discarded by other labeling methods. We define labeler-ability (α) and instance-difficulty (β) as:
$\alpha_j = (p_j - q_j)(1 - E_j), \quad E_j = -p_j log(p_j) - q_j log(q_j)$
and $p_j = \frac{n_j^+}{n}$, $q_j = 1 - p_j$ and n_j^+ is the number of correctly labeled instances from $\mathcal{D}'$ by labeler j.
$\beta_i = (p'_i - q'_i)(1 - E'_i) + 1, \quad E'_i = -p'_i log(p'_i) - q'_i log(q'_i)$
where $p'_i = \frac{M_i^+}{M}$, $q'_i = 1 - p'_i$, p'_i is the probability of getting a correct label for instance i, from the crowd labeler and M_i^+ is the number of correct labels given to the instance i. All these values are calculated using the expert labeled instances. Then α, β are used for label aggregation as follows:
$A_i = sign(\sum_{j=1}^{M}\sigma(|c\alpha_j\beta_i|) * L_{ij} * sign(\alpha_j\beta_i))$
The βs for the rest of the data are estimated using equation 1. Here c is the scaling factor and $sign(\alpha\beta)$ is used to flip the label provided by the malicious labeler i.e., when α is negative. Both versions of ELICE have a cluster-based variant in which rather than making a random choice of instances from the whole dataset, clusters of data are first formed using any clustering approach e.g., K-means.
The motivation behind developing the third version of ELICE [Khattak and Salleb-Aouissi, 2016] was to further improve the accuracy by using the crowd-labels, which unlike expert-labels, are available for the whole dataset and may provide a more comprehensive view of the labeler ability and instance

difficulty. This is especially helpful for the case when the domain experts do not agree on one label and ground truth is not known for certain. Therefore, incorporating more information beyond expert-labels can provide better results. Besides taking advantage of expert-labeled instances, the third version of ELICE, incorporates pairwise/circular comparison of labelers to labelers and instances to instances. In this variant of ELICE, we use a generalization of the model in [Bradley and Terry, 1952; Huang *et al.*, 2006]. We show empirically that our approaches are robust even in the presence of a large proportion of low-quality labelers in the crowd (Figure 1). Furthermore, we derive a lower bound of the number of expert labels needed [Khattak and Salleb-Aouissi, 2013].

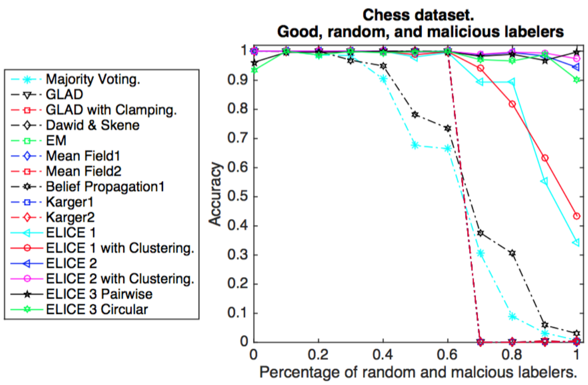

Figure 1: UCI Chess Dataset [Asuncion and Newman, 2007]. Accuracy of Majority voting, GLAD (with and without clamping) [Whitehill *et al.*, 2009], Majority voting, Dawid and Skene [Dawid and Skene, 1979], EM (Expectation Maximization), Karger's iterative method [Karger *et al.*, 2014], Mean Field algorithm and BP [Liu *et al.*, 2012] and ELICE (all versions and variants) with 20 expert-labeled instances. Good labelers: 0-35% mistakes, Random labelers: 35-65% mistakes, Malicious labelers: 65-100% mistakes. Accuracy vs. percentage of random and malicious labelers averaged over 50 runs. We start with all good labelers and keep on increasing the percentage of random and malicious labelers.

3 Bayesian Approach

Currently, we are exploring Bayesian method for parameter estimation. Our new approach [Khattak and Salleb-Aouissi, 2015] is inspired by Item Response Theory (IRT) [Lord, 1952]. IRT aims to design and analyze test scoring strategies by modeling student ability, question difficulty, question clarity and probability of correctness of the answer to the question. Similarly in crowd-labeling, we model labeler ability, instance difficulty, clarity of the question about the instance and probability of correctness of label. The crowd-labeling scenario is more challenging, as unlike IRT model, the parameters as well as final labels are unknown. To deal with this challenge, we use expert-labeled instance (ground truth) for a small percentage of data to learn the parameters. These parameters are used for aggregation of multiple crowd-labels for the rest of the dataset with no ground truth available. Our new model is as follows:
$P[c|y_{ij} = c, \gamma_c, \beta_i, \delta_i, \pi_c^{(j)}] = [\text{logit}^{-1}(\delta_i(\gamma_c + \pi_c^{(j)} - \beta_i))]$
where $c \in \{-1, 1\}$: class/category, y_{ij}: Label provided by labeler j to instance i, $\pi_c^{(j)}$: per-class ability of labeler j, β_i : difficulty of instance i, γ_c: prevalence of class c, δ_i : clarity of question asked about instance i. Experiments are ongoing.

4 Conclusion

We propose a set of methodologies to advance the state-of-the-art in crowd-labeling methods using a handful expert-labeled instances. Our future plans include developing methodologies for analyzing and modeling labeler's variable performance due to fatigue, stress and boredom. We hope that it will help in further improving crowd-labeling accuracy.

References

[Asuncion and Newman, 2007] A. Asuncion and D.J Newman. UCI machine learning repository. In *University of California, Irvine. School of Information and Computer Sciences*, 2007.

[Bradley and Terry, 1952] Ralph Allan Bradley and Milton Terry. Ranking analysis of incomplete block design: I. the method of paired comparisons. In *Biometrika*, pages 324–345, 1952.

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

[Huang *et al.*, 2006] Tzu-Kuo Huang, Chih-Jen Lin, and Ruby C. Weng. Ranking individuals by group comparisons. In *International Conference on Machine Learning (ICML)*, ICML '06, pages 425–432, New York, NY, USA, 2006.

[Karger *et al.*, 2014] David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

[Khattak and Salleb-Aouissi, 2011] Faiza Khan Khattak and Ansaf Salleb-Aouissi. Quality control of crowd labeling through expert evaluation. In *NIPS 2nd Workshop on Computational Social Science and the Wisdom of Crowds, Granada, Spain.*, 2011.

[Khattak and Salleb-Aouissi, 2013] Faiza Khan Khattak and Ansaf Salleb-Aouissi. Robust crowd labeling using little expertise. In *Sixteenth International Conference on Discovery Science, Singapore*, 2013.

[Khattak and Salleb-Aouissi, 2015] Faiza K. Khattak and Ansaf Salleb-Aouissi. An item response theory (IRT) like approach to crowd-labeling. In *Workshop for Women in Machine Learning (WiML 2015) held in conjunction with NIPS, Montreal, Canada. 2015.*, 2015.

[Khattak and Salleb-Aouissi, 2016] Faiza K. Khattak and Ansaf Salleb-Aouissi. *Robust Crowd Labeling using Expert evaluation and Pairwise Comparison*. Submitted to Journal of Artificial Intelligence. Special Track on Human Computation and AI, 2016.

[Liu *et al.*, 2012] Qiang Liu, Jian Peng, and Alex Ihler. Variational inference for crowdsourcing. In *NIPS*, 2012.

[Lord, 1952] F. Lord. *A Theory of Test Scores*. Psychometric Monograph No. 7, 1952.

[Whitehill *et al.*, 2009] Jacob Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, pages 2035–2043, 2009.