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

EM-Based Inference of True Labels Using Confidence Judgments[*]

Satoshi Oyama
Hokkaido University
oyama@ist.hokudai.ac.jp

Yukino Baba
The University of Tokyo
yukino_baba@mist.i.u-tokyo.ac.jp

Yuko Sakurai
Kyushu University
ysakurai@inf.kyushu-u.ac.jp

Hisashi Kashima
The University of Tokyo
kashima@mist.i.u-tokyo.ac.jp

Abstract

We have developed a method for accurately inferring true labels from labels provided by crowdsourcing workers, with the aid of self-reported confidence judgments in their labels. Although confidence judgments can be useful information for estimating the quality of the provided labels, some workers are overconfident about the quality of their labels while others are underconfident. To address this problem, we extended the Dawid-Skene model and created a probabilistic model that considers the differences among workers in their accuracy of confidence judgments. Results of experiments using actual crowdsourced data showed that incorporating workers' confidence judgments can improve the accuracy of inferred labels.

Introduction

An inherent problem in applying crowdsourcing is *quality control*. In contrast with well-controlled cases with reliable, screened workers, labels provided by crowdsourcing workers tend to contain many errors due to their varied abilities and dedication levels. Several methods have been proposed for inferring true labels from worker provided labels that consider the differences in the abilities of workers to provide true labels. In the most well-known method, proposed by Dawid and Skene (1979), each worker is assumed to have a distinct conditional probability of producing his/her label given a (an unknown) true label. Several other methods also consider the difficulty of the task as well as the ability of the workers in inferring the true labels (Whitehill et al. 2009; Welinder et al. 2010).

The studies mentioned above took a machine-based approach: the label or worker quality is automatically estimated using a statistical inference or machine learning technique. In contrast, we use a human-based approach to determining label quality: the workers are directly asked to report their level of confidence in the labels they provide. Since a worker can easily judge the difficulty of a task and his/her ability to perform it, he/she is the person best suited to evaluate the quality of the label given.

The possibility of using these confidence judgments to improve the quality of crowdsourced labels was investigated by a few researchers (Ipeirotis 2009; Kazai 2011). Confidence judgments given by workers should be useful information for inferring the true labels. For example, if a worker's confidence about his/her label for an item is high, the likelihood that his/her label coincides with the true label is high. However, a way to effectively incorporate them in an inference algorithm has not been established. In addition, the quality of the confidence judgments varies among workers just as the label quality does. Some workers may be overconfident and report a high level of confidence even though their labels are actually incorrect, while other workers may be underconfident and report a low level of confidence even though their labels are actually correct. Some workers may be quite accurate in judging their actual abilities, i.e., they are "well-calibrated."

In this work, we assume that each worker has a distinct conditional distribution for the confidence judgments given the true label and his/her labels. This enables us to model each worker's particular tendency in giving confidence judgments, such as an overconfident worker who gives a high level of confidence with high probability even when the true label and his/her label are different or an underconfident worker who gives a low level of confidence with high probability even when the true label and his/her label are the same.

Proposed Model

The problem setting is similar to that of Dawid and Skene (1979). There are N data items and J crowdsourcing workers (each worker does not necessarily label all items). Let $\mathcal{J}_i \subseteq \{1, \ldots, J\}$ be the subset of workers who labeled item i. $t_i \in \{0, 1\}$ $(i \in \{1, \ldots, N\})$ is the true label for data item i, and $y_{ij} \in \{0, 1\}$ $(j \in \mathcal{J}_i)$ is the label for data i given by worker j. In contrast to the setting of Dawid and Skene (1979), we collect additional information from workers as well as the label estimates. Each worker is asked to assign a confidence score to his/her labels. The level of confidence of worker j in his/her label for item i is given by $c_{ij} \in \{0, 1\}$ $(j \in \mathcal{J}_i)$. If the worker is confident, $c_{ij} = 1$; otherwise, $c_{ij} = 0$. The confidence score is given as a binary variable for simplicity, but the model can be easily extended to enable the use of more general confidence scores, such as multi-level scales and numerical scores.

[*] The full version of this paper appeared in the Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI) (Oyama et al. 2013).

We propose using a probabilistic generative model of the confidence judgments as well as the labels given by crowdsourcing workers. With this model, we can use workers' confidence judgments as well as their labels to infer the value of the true labels. Our models are given as a factorization of the joint distribution:

$$\begin{aligned} p(\{t_i\}, \{y_{ij}\}, \{c_{ij}\}) \\ = \prod_{i \in \{1,\ldots,N\}} \prod_{j \in \mathcal{J}_i} p(c_{ij}|y_{ij}, t_i) p(y_{ij}|t_i) p(t_i). \end{aligned}$$

The value of a true label for item i takes 1 with probability p_i and 0 with probability $1 - p_i$; that is, it is sampled from a Bernoulli distribution with parameter p_i.

Worker labels $\{y_{ij}|j \in \mathcal{J}_i\}$ for item i are conditionally independent given true label t_i. $\boldsymbol{\alpha}^{(j)} = \{\alpha_0^{(j)}, \alpha_1^{(j)}\}$ represents the set of parameters for worker j, where $\alpha_0^{(j)}$ is the probability of worker j giving label 1 if the true label is 0, and $\alpha_1^{(j)}$ is the probability of worker j giving label 1 if the true label is 1. Therefore, when $t_i = 1$, label y_{ij} given by worker j for item i is sampled from a Bernoulli distribution with parameter $\alpha_1^{(j)}$. Similarly, when $t_i = 0$, label y_{ij} given by worker j for item i is sampled from a Bernoulli distribution with parameter $\alpha_0^{(j)}$.

Worker j's confidence judgment c_{ij} for his/her label for item i depends on the true label t_i and his/her label y_{ij}, and it is also sampled from a Bernoulli distribution. In the proposed model, $\boldsymbol{\beta}^{(j)} = \{\beta_{00}^{(j)}, \beta_{01}^{(j)}, \beta_{10}^{(j)}, \beta_{11}^{(j)}\}$ is the set of parameters specific to worker j. Here, for example, $\beta_{00}^{(j)}$ is the probability that worker j's confidence $c_{ij} = 1$ when true label $t_i = 0$ and worker j's label $y_{ij} = 0$. In this case, the confidence is sampled from the following distribution: $p(c_{ij}|t_i = 0, y_{ij} = 0) = (\beta_{00}^{(j)})^{c_{ij}}(1 - \beta_{00}^{(j)})^{(1-c_{ij})}$. When $t_i = 0$ and $y_{ij} = 1$, the confidence is sampled from the following distribution: $p(c_{ij}|t_i = 0, y_{ij} = 1) = (\beta_{01}^{(j)})^{c_{ij}}(1 - \beta_{01}^{(j)})^{(1-c_{ij})}$. The conditional distributions for the other two cases, $p(c_{ij}|t_i = 1, y_{ij} = 0)$ and $p(c_{ij}|t_i = 1, y_{ij} = 1)$ are similarly defined.

Our goal is to infer the set of true labels $\{t_i\}$ given the set of workers' labels $\{y_{ij}\}$ and the set of confidence judgments $\{c_{ij}\}$. Similar to the approach of Dawid and Skene (1979), we use the EM algorithm to obtain the maximum likelihood estimate of model parameters $\{\boldsymbol{\alpha}^{(j)}\}$ and $\{\boldsymbol{\beta}^{(j)}\}$, with true labels $\{t_i\}$ as latent variables. The EM algorithm for the proposed model alternately performs two steps until convergence.

E-step: Estimate the expected values of unobserved variables $\{t_i\}$ by using the current estimates of parameters $\{\boldsymbol{\alpha}^{(j)}\}$ and $\{\boldsymbol{\beta}^{(j)}\}$.

M-step: Estimate parameters $\{\boldsymbol{\alpha}^{(j)}\}$ and $\{\boldsymbol{\beta}^{(j)}\}$ by using the current expectations of $\{t_i\}$.

Experiments

To evaluate the effectiveness of using confidence judgments in inferring true labels, we conducted experiments using Amazon Mechanical Turk. We chose ten images from the Caltech-UCSD Birds 200 dataset and asked crowdsourcing workers to choose one of two bird names as the label for each image. We also asked them to report their level of confidence in each choice. We asked 100 workers to label the same ten images. To see the effect of the number of workers per task on accuracy, we split the workers into groups of equal size, inferred the true labels from the worker labels and confidence judgments within each group, and averaged the accuracies of the true labels obtained from each group. We conducted experiments with four different group sizes: 5, 10, 20, and 50. The average accuracies for each group size were obtained with majority vote, the Dawid-Skene model, the proposed model. As shown in Figure 1, with 50 workers, all three models and even a simple majority vote provided sufficient accuracy due to the high level of redundancy. In practice, however, the number of workers that can be used for a task is limited due to cost. When the number of workers was 5 or 10, the the model using the confidence judgments achieved better accuracy than majority vote and the Dawid-Skene model.

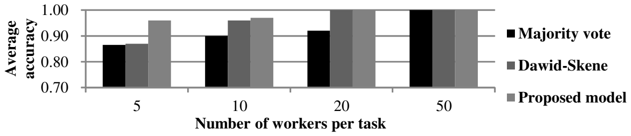

Figure 1: Results for image labeling

We also conducted experiments using another dataset, one containing 120 binary questions on general knowledge. See Oyama et al. (2013).

Acknowledgments

SO was supported by JSPS KAKENHI 24650061. YB and HK were supported by the FIRST program. YS was supported by the JST PRESTO program.

References

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.

Ipeirotis, P. 2009. How good are you, Turker? http://www.behind-the-enemy-lines.com/2009/01/how-good-are-you-turker.html.

Kazai, G. 2011. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *ECIR*.

Oyama, S.; Baba, Y.; Sakurai, Y.; and Kashima, H. 2013. Accurate Integration of Crowdsourced Labels Using Workers' Self-reported Confidence Scores. In *IJCAI*.

Welinder, P.; Branson, S.; Belongie, S.; and Perona, P. 2010. The Multidimensional Wisdom of Crowds. In *NIPS 23*.

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 *NIPS 22*.