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

A Ground Truth Inference Model for Ordinal Crowd-Sourced Labels Using Hard Assignment Expectation Maximization

Siamak Faridani, Georg Buscher
Microsoft
One Microsoft Way
Redmond, WA

Ya Xu*
LinkedIn
2029 Stierlin Court
Mountain View, CA

Abstract

In this paper we propose an iterative approach for inferring a ground truth value of an item from judgments collected form online workers. The method is specifically designed for cases in which the collected labels are ordinal. Our algorithm works by iteratively solving a hard-assignment EM model and later calculating one final expected value after the convergence of the EM procedure.

Introduction

We focus on the problem of inferring the ground truth when ordinal labels are collected from online judges (also known as Turkers). Traditionally, when the labels do not have an order, the EM model by Dawid and Skene (1979) is used for inferring the ground truth. When labels are ordinal, various ad-hoc methods like median, mean, or robust mean or median are used by practitioners. Our algorithm can be considered as an extension of the model by Smyth et al. (1995) for ordinal labels.

Problem Statement

We assume *m* items are classified into *n* different ordinal categories by online judges. The number of judgments that we collect on each item can vary from item to item and is not fixed. Additionally, we assume that each item can be assigned to one and only one of the categories. However the true label of the item might be in between two labels. For example, in the case of measuring the relevance of a machine-generated caption to a document, the inputs might be on the scale of ["very good", "good", "neutral", "bad", "very bad"]. The final inferred ground truth by our algorithm, instead of "good" or "very good" can be "0.7good+0.3very good" suggesting the actual latent label is closer to "good" than "very good".

We assume for item *i* we collect a judgment from judge *j* and the judge indicates that the item is from category *k*. Denote $X_{ik}^{(j)}$ as the indicator function that captures this judgment and is defined as 1 if item *i* is classified as class *k* by judge *j* and 0 otherwise. Let's consider another identity variable T_{ik} that is 1 if ground truth value for *i* is closest to category *k* and 0 otherwise. So for each item *i* and class *k* we want to calculate $P(T_{ik} \mid data_i)$. In which $data_i$ is the collection of judgments that we have collected for item *i*. After finding all the values of $P(T_{ik} \mid data_i)$ by fixing *i* and varying *k* we perform our final round of inference by calculating the expected value of the label $E(P(T_{ik} \mid data_i))$.

I. Calculating the Posterior

For calculating the values of $P(T_{ik} \mid data_i)$ we can use the Bayes' rule

$$P(T_{ik} \mid data_i) \xrightarrow{Bayes' \ Rule} \frac{P(\ data_i \mid T_{ik})P(T_{ik})}{P(\ data_i)}$$

$P(\ data_i)$ is easy to find knowing that it is the joint probability of individual judgments that can be separated if we assume that judges are not influenced by each other.

$$P(\ data_i) = \prod_{j=\{judges\ that\ judged\ i\}} P(data_i^j)$$

$data_i^j$ is the judgment from judge *j* on item *i*. This is the bias for each judge and easy to calculate from historical data. Any implementation of a naïve Bayes' model needs to use a Laplace smoothing to prevent divergence. Similarly, $P(\ data_i \mid T_{ik})$ can be broken into the following

$$P(\ data_i \mid T_{ik}) = \prod_{j=\{judges\ that\ judged\ i\}} P\left(data_i^j \middle| T_{ik}\right)$$

This is a conditional probability for each judge and can be found by conditioning on the ground truth. These are the elements of the confusion matrix. The MLE of $P(T_{ik})$ is in fact done by counting. See Manning et al. (2008) for more. $P(T_{ik})$ is the number of judgments on *i* that are from label *k* over the total number of judgments on *i* regardless of *k*.

* This work was done when the author was at Microsoft.

We need to add 1 to the numerator and add length(k) to the denominator for Laplace smoothing.

$$P(T_{ik}) = \frac{\#\ of\ judges\ that\ labeled\ item\ i\ to\ be\ label\ k{+}1}{total\ number\ of\ judgments\ on\ item\ i{+}length(k)}$$

So at this point we can calculate the value of $P(T_{ik} \mid data_i)$ for any i and k as below:

$$P(T_{ik} \mid data_i) = P(T_{ik}) \prod_{j=\{judges\ that\ judged\ i\}} \frac{P(data_i^j|T_{ik})}{P(data_i^j)}$$

From this point on we can remove the denominator and solve the following argmax on the log likelihood if we only had categorical labels without any order.

$$k = argmax_k \left[\ln(P(T_{ik})) \prod_{j=\{judges\ that\ judged\ i\}} \ln(P(data_i^j|T_{ik})) \right]$$

II. Calculating The Expected VALUE

The expected score is easy to calculate from the likelihood $E(P(T_{ik} \mid data_i)) =$

$$\sum_{k\,\in\{categories\}} k.P(T_{ik}) \prod_{j=\{judges\ that\ judged\ i\}} \frac{P(data_i^j|T_{ik})}{P(data_i^j)}$$

Algorithm

We propose the following algorithm for the cases in which the collected labels are ordinal:

Step 1. Find an initial approximation of $P(T_{ik} \mid data_i)$ by using mean, median, or mean and median after outlier removal
Step 2. Update $P(data_i^j|T_{ik})$
Step 3. For every i find $P(T_{ik} \mid data_i)$
Step 4. If any of T_{ik} values changed, go to *step 2*
Step 5. Calculate $E(P(T_{ik} \mid data_i))$

Evaluation

Here, we provide a preliminary evaluation of the model, but we are planning to run larger experiments in the near future. For this model we ran two small blind experiments. For the first experiment, we took 5000 documents, each of which with one machine-generated title. For each title, 5 judges had provided relevance labels from the set of ["excellent", "good", "fair", "poor", "embarrassing"]. We ran both GLAD (Whitehill et al. 2009) and our algorithm on this set. Since GLAD only provides the most likely label, for a fair comparison we round the output of our algorithm to the closest label on the number line (assuming the distance between each two consecutive label is 1). We randomly took 30 items for which our algorithm and GLAD disagreed. The results of the two algorithms were anonymized and sent to a summarization expert to evaluate which one is closer to the expert label. The outcome of this was 18 to 12 in favor of our algorithm. We also noted that GLAD sometimes comes up with labels that are very far from the collected judgments. In an additional short experiment we asked 3 experts to compare 18 pairs of search engine result pages for 18 different queries and label them as ["Left Much Better", "Left Better", "Right Better", "Right Much Better"]. We later compared their results with the results of our algorithm, GLAD, and median. In 12 cases (67%) our algorithm was the closest to the expert label, in 10 cases (55%) median was the closest model and in 7 cases (38%) GLAD produced the closest label.

Conclusion and Future Work

This paper extends the previous work on aggregating categorical judgments to ordinal labels. Our algorithm does not require any parameter tuning and can serve as turnkey algorithm for aggregating categorical and ordinal judgments. It works by iteratively solving a hard-assignment EM model which is reduced to solving a set of naïve Bayes' steps. At the end we calculate the expected value of $E(P(T_{ik} \mid data_i))$ which can be looked at as a weighted mean of judgments where the weight of each judgment is represented by the reputation of the corresponding judge. This reputation is taken from the error rates (confusion matrix) for each judge on each label. The next step for us is to compare this algorithm with mean, median and GLAD on a larger set of data and characterize their behavior based on the type of task, number of judgments on each item, and the number of allowable labels.

References

Dawid, A. P., & Skene, A. M. (1979). *Maximum likelihood estimation of observer error-rates using the EM algorithm*. Applied Statistics, 20-28.

Smyth, Padhraic, et al. "*Inferring ground truth from subjective labelling of venus images.*" Advances in neural information processing systems (1995): 1085-1092.

Whitehill, J., Wu, T. F., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). *Whose vote should count more: Optimal integration of labels from labelers of unknown expertise*. In Advances in neural information processing systems (pp. 2035-2043).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval (Vol. 1)*. Cambridge: Cambridge University Press.