COMBINING SOFT DECISIONS OF SEVERAL UNRELIABLE EXPERTS

Jacob Goldberger

Engineering Faculty, Bar-Ilan University, Israel
jacob.goldberger@biu.ac.il

ABSTRACT

In this study we address the problem of integrating information from several experts with unknown levels of expertise. In the usual setup each expert expresses her opinion by choosing one of the options. Here we assume that each expert provides her opinion in a soft manner via a distribution on the possible options. The goal is to find the reliability level of each expert and to optimally integrate their information. We develop an estimation algorithm which is an instance of the EM algorithm and an efficiently computed approximation of the E-step. Finally we present simulations that demonstrate the improved performance of the proposed approach.

***Index Terms*—** Unreliable experts, EM algorithm, crowdsourcing

1. INTRODUCTION

In this study we address the problem of integrating information from several experts with unknown levels of expertise. In the standard setup there is a set of questions each being associated with an unobserved correct answer. We direct these questions to a set of experts. When a question is assigned to an expert, the answer we get may be inaccurate depending on his or her level of expertise. In this case the challenge is to find the correct answer and, as a by-product, to assess the reliability of each expert. This problem is closely related to monitoring crowdsourcing systems such as Amazon Mechanical Turk. Crowdsourcing is an effective paradigm for human-powered problem solving which is now in widespread use for large-scale data-processing tasks [5][4].

A principled way to address this problem is to build generative probabilistic models for the expert decision processes, and assign labels using standard inference tools. The expert reliability is viewed as an unknown parameter. A line of works applied the EM algorithm to this task incorporating either simple or more complicated generative models (e.g. [8][2][7][10][1][9][6]). Another line of works applied spectral methods to this problem [3].

In this study we address a more general instance where the experts do not provide an explicit answer or, equivalently, we do not know the expert's exact opinion. Instead, each expert splits his vote among the possible answers. The opinion of an expert is thus provided in the form of a distribution over the possible answers. This situation occurs naturally when the expert is an automatic probabilistic classifier such as a logistic regression or a neural network.

In this work we first generalize the classical EM algorithm to the case where even the so-called observed random variable is not completely observed. Then we utilize this EM extension to our problem of integrating a set of unreliable soft experts. Finally we apply our approach to simulated data and show its improved performance.

2. COMBINING SEVERAL EXPERTS

In this study we focus on the problem of combining opinions from several unreliable experts where the opinion is in the form of a distribution on the possible options. We start by reviewing the EM based approach for the simpler and standard case where the experts provide explicit answers. Assume $x_1, ..., x_n$ are random variables that are uniformly sampled from a finite set A. The values of $x_1, ..., x_n$ are not directly observed. Instead, there is a set of m 'experts' and the opinion of expert i on the value x_j is denoted by $y_{ij} \in A$. We assume that each expert i is associated with a reliability probability θ_i of providing the correct answer. To simplify the modeling we further assume that when the expert makes a wrong decision he samples uniformly from the $|A| - 1$ alternatives:

$$p(y_{ij}|x_j = a; \theta_i) = \begin{cases} \theta_i & \text{if} \quad y_{ij} = a \\ \\ \frac{1-\theta_i}{|A|-1} & \text{if} \quad y_{ij} \neq a \end{cases}, \qquad a \in A \tag{1}$$

Let $y_j = \{y_{1j}, ..., y_{mj}\}$ be the opinions, independently provided by the m experts, on the value of x_j.

$$p(y_j|x_j = a; \theta) = \prod_{i=1}^{m} p(y_{ij}|x_j = a; \theta_i) \tag{2}$$

such that $\theta = \{\theta_1, ..., \theta_m\}$. Given the experts' opinions we can compute the posterior distribution of the possible values of x_j. Applying Bayes' rule, we obtain:

$$p(x_j = a|y_j; \theta) = \frac{p(y_j|x_j = a; \theta)}{\sum_{b \in A} p(y_j|x_j = b; \theta)}, \qquad a \in A \tag{3}$$

and from that we can also extract a hard decision:

$$\hat{x}_j = \arg\max_{a \in A} p(x_j = a | y_j; \theta). \tag{4}$$

If all the experts have the same reliability, then $\hat{x}_j$ is simply a majority voting decision.

In the case where the reliability parameters are unknown, our goal is to find them (and the value of the unobserved r.v.) using the given expert information set (y_{ij}). The log-likelihood function is:

$$\begin{aligned} L(\theta) &= \sum_{j=1}^{n} \log p(y_j; \theta) \\ &= \sum_{j=1}^{n} \log(\frac{1}{|A|} \sum_{a \in A} p(y_j | x_j = a; \theta)). \end{aligned} \tag{5}$$

Since $x_1, ..., x_n$ are hidden variables and $\theta_1, ..., \theta_m$ are unknown parameters, we utilize the EM algorithm to find the maximum-likelihood parameters. The EM algorithm handles the parameter estimation task by iterating between the E and M steps. The E-step is:

$$w_j(a) = p(x_j = a | y_j; \theta_0) = \frac{p(y_j | x_j = a; \theta_0)}{\sum_{b \in A} p(y_j | x_j = b; \theta_0)} \tag{6}$$

$$j = 1, ..., n, \quad a \in A$$

where θ_0 is the current values of the parameter-set. The M-step is:

$$\theta_i = \frac{1}{n} \sum_{j=1}^{n} w_j(y_{ij}), \qquad i = 1, ..., m \tag{7}$$

The updated θ_i is the expected number of the times that expert i provided the correct decision. The EM algorithm is notoriously known to get stuck in a local maximum point of the likelihood function. Hence, it is important to choose a meaningful way to initialize the model's parameters. A reasonable initialization of the EM algorithm is setting all reliability parameters to have the same value (e.g. $\theta_i = 0.7$ for all $i = 1, ..., m$).

3. EXTENSION OF THE EM ALGORITHM TO SOFT OBSERVATION VALUES

In this section we introduce an extension of the classical EM algorithm that can handle soft observed information. We start with a brief description of the EM algorithm. Assume y is an observed sample from a random variable Y whose density function is $p(y; \theta)$ such that θ is an unknown parameter. The goal of the EM algorithm is to facilitate the maximum likelihood parameter estimation:

$$\theta_{ML} = \arg\max_{\theta} \log p(y; \theta) \tag{8}$$

by introducing so-called a hidden random variable, denoted by X, and by introducing a joint distribution $p(x, y; \theta)$ such that the given marginal distribution $p(y; \theta)$ satisfies $p(y; \theta) = \sum_x p(x, y; \theta)$. The EM algorithm is an iterative procedure where each iteration is composed of two steps. In the E-step we compute the following conditional expectation:

$$\begin{aligned} Q(\theta, \theta_t) &= \sum_x p(x | y; \theta_t) \log p(x, y; \theta) \\ &= E_{p(x|y;\theta_t)}(\log p(x, y; \theta) | Y = y) \end{aligned} \tag{9}$$

where θ_t is the current estimation of the parameter θ. In the M-step we compute an updated parameter estimation by solving the following maximization problem:

$$\theta_{t+1} = \arg\max_{\theta} Q(\theta, \theta_t). \tag{10}$$

The success of the EM algorithm is based on the fact that in many cases solving the maximization problem (10) is much easier than directly solving the maximum-likelihood problem (8). A major feature of the EM algorithm is a monotone increase of the likelihood function, i.e., $p(y; \theta_t) \leq p(y; \theta_{t+1})$.

Assume now that y is not fully observed. Instead we are only given a distribution $q(y)$ on the possible values that the r.v. Y can obtain. We can view this situation as though we only receive a noisy (soft) information about the sampled value of the r.v. y. We still want to estimate the unknown parameter θ and the value of the hidden r.v. x. The cost function we aim to optimize is:

$$\begin{aligned} \theta_{MKL} &= \arg\max_{\theta} \sum_y q(y) \log p(y; \theta) \\ &= \arg\min_{\theta} KL(q(y) || p(y; \theta)) \end{aligned} \tag{11}$$

where KL is the Kullback-Leibler divergence and θ_{MKL} is the minimum KL parameter estimation. Note that if y is completely known (i.e. $q(y)$ is a delta distribution) then this cost function coincides with the likelihood function that is maximized by the standard EM algorithm. Maximum likelihood estimation is known to be identical to the minimization of KL divergence between the empirical distribution and the model distribution. Our cost function for soft observations (11) follows this interpretation.

We next extend the EM algorithm to the soft-observation situation where instead of observing y, we are only given a distribution $q(y)$. We dub this EM extension as the Soft Observation EM (SOEM) algorithm. It can be easily verified that:

$$\sum_y q(y) \log p(y; \theta) = \sum_{y,x} q(y) p(x | y; \theta) \log \frac{p(x, y; \theta)}{p(x | y; \theta)}. \tag{12}$$

The concavity of the $\log()$ function (or equivalently the Jensen inequality) implies that for every conditional distribution $r(x|y)$:

$$\sum_y q(y)\log p(y;\theta) \geq \sum_{y,x} q(y)r(x|y)\log\frac{p(x,y;\theta)}{r(x|y)}. \tag{13}$$

Combining Eq. (12) and Eq. (13) we obtain:

$$\begin{aligned}\theta_{MKL} &= \arg\max_\theta \max_r \sum_{y,x} q(y)r(x|y)\log\frac{p(x,y;\theta)}{r(x|y)} \\ &= \arg\min_\theta \min_r KL(r(x|y)q(y)||p(x,y;\theta)).\end{aligned} \tag{14}$$

Approximating the double minimization in Eq. (14) with an alternating minimization we get the SOEM algorithm. In the E-step we compute the following conditional expectation.

$$\begin{aligned}Q(\theta,\theta_t) &= \sum_y q(y)\sum_x p(x|y;\theta_t)\log p(x,y;\theta) \\ &= E_{q(y)p(x|y;\theta_t)}(\log p(x,y;\theta)).\end{aligned} \tag{15}$$

Note that unlike the standard EM where the expectation is based on the conditional distribution $p(x|y;\theta_t)$, here the expectation is performed based on the joint distribution $q(y)p(x|y;\theta_t)$. The M-step of the SOEM remains the same: $\theta_{t+1} = \arg\max_\theta Q(\theta,\theta_t)$. The alternating minimization view of the SOEM algorithm implies that it monotonically improves the estimated parameter θ in the sense that:

$$KL(q(y)||p(y;\theta_t)) \geq KL(q(y)||p(y;\theta_{t+1})).$$

Once we have found the ML estimation of the model parameters θ we can reconstruct the hidden random variable x:

$$\hat{p}(x) = \sum_y q(y)p(x|y;\theta). \tag{16}$$

4. A SOFT VERSION OF EXPERT OPINIONS

We next extend the problem of combining the opinions of several experts to the case where the experts provide their opinions in the form of a distribution over the set of possible decisions A. Assume there are m experts who provide opinions on the values of the $x_1,...x_n$ which are not directly observed. The experts do not provide hard decisions. Instead, each expert splits his vote among the possible $|A|$ values. For example, assume that y is obtained as an output of a probabilistic classifier such as a logistic regression or a neural network. The opinion of expert i on the value of x_j is thus provided in the form of a distribution:

$$q_{ij}(b) = p(y_{ij} = b), \qquad b \in A \tag{17}$$

Assuming the experts' opinions are independently generated, we use the following notation for the soft opinions on the values of x_j:

$$q_j(a) = q_j(a_1,...,a_m) = \prod_{i=1}^m q_{ij}(a_i), \qquad a = (a_1,...,a_m) \in A^m \tag{18}$$

Following the definition of the SOEM algorithm in the previous section, the cost function $L(\theta)$ we want to optimize is:

$$\sum_{j=1}^n \sum_{a\in A^m} q_j(a)\log(\frac{1}{|A|}\sum_{b\in A} p(y_j = a|x_j = b;\theta)). \tag{19}$$

The optimal parameter, therefore, is:

$$\theta_{MKL} = \arg\max_\theta L(\theta) = \arg\min_\theta \sum_{j=1}^n KL(q_j||p(y_j;\theta)). \tag{20}$$

The optimal parameter can be found by applying the SOEM algorithm defined above. The auxiliary function for this case is:

$$Q(\theta,\theta_0) = \sum_{j=1}^n \sum_{a\in A^m} q_j(a)\sum_{b\in A} p(x_j = b|y_j = a;\theta_0)\times$$

$$\sum_{i=1}^m (1_{\{a_i=b\}}\log\theta_i + 1_{\{a_i\neq b\}}\log(\frac{1-\theta_i}{|A|-1}))$$

where the term $p(x_j = b|y_j = a;\theta_0)$ is computed using Eq. (3). The E-step is:

$$w_j(i) = \sum_{a\in A^m} q_j(a)p(x_j = a_i|y_j = a;\theta_0) \tag{21}$$

$$j = 1,...,n, \quad i = 1,...,m$$

$w_j(i)$ is the posterior probability that expert 'i' provided the correct decision on the value of x_j. The M-step is:

$$\theta_i = \frac{1}{n}\sum_{j=1}^n w_j(i), \qquad i = 1,...,m. \tag{22}$$

Note that the complexity of computing the expressions $w_j(i)$ in the E-step is exponential in the number of experts. Hence, in the case where there are many experts or the set A is large, a direct computation of the E-step is not feasible. We can approximate the expectation $w_j(i) = E_{\{q_j(a)\}}p(x_j = a_i|y_j = a;\theta_0)$ by sampling $a^1,...,a^k \in A^m$ from the distribution q_j. Note that sampling from the distribution q_j is easy since it is a joint distribution of independent r.v. (see Eq. (18)). Using the law of large numbers, we can replace the expectation by a sample average:

$$w_j(i) \approx \frac{1}{k}\sum_{t=1}^k p(x_j = a_i^t|y_j = a^t;\theta_0). \tag{23}$$

Once we have found the model parameter-set θ we can compute the posterior distribution of x_j based on the experts' opinions.

$$p(x_j = b|y_j;\theta) = \sum_{a\in A^m} q_j(a)p(x_j = b|y_j = a;\theta), \qquad b \in A \tag{24}$$

In the case where an exact computation of the posterior distribution (24) in not feasible, we can use sampling methods, similar to the one described above to approximately compute the posterior distribution. Finally, the hard-decision prediction is:

$$\hat{x}_j = \arg\max_{b \in A} p(x_j = b | y_j; \theta). \tag{25}$$

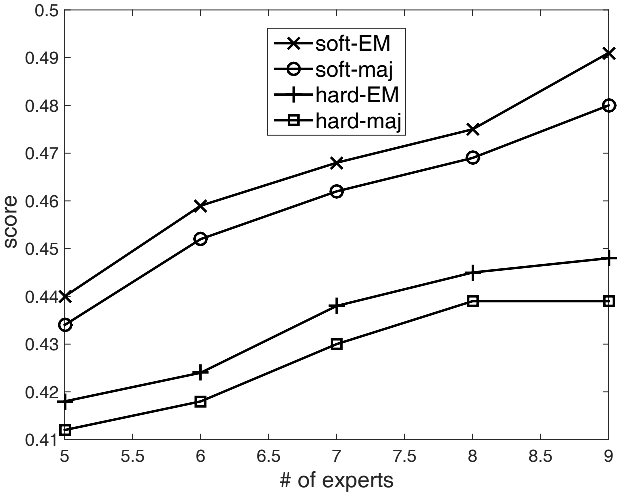

Fig. 1. Algorithm performance as a function of the number of experts.

5. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method for combining soft expert decisions we conducted the following simulation. We set the label-set A to be $\{0, 1, 2\}$. We uniformly sampled $n = 200$ 'correct' labels $x_1, ..., x_n \in A$. For each expert i we chose a reliability parameter θ_i by sampling θ_i uniformly from the interval $[0.4, 0.7]$. There are many ways to simulate a soft decision procedure. We used the following scheme: for each object item j and expert i, we first sampled a multinomial distribution U_{ij} from the uniform distribution over the standard $(|A| - 1)$-simplex (a.k.a. the flat Dirichlet distribution). We next sampled an expert opinion y_{ij} using Eq. (1). Then the soft-decision of expert i on item j is obtained as:

$$q_{ij}(a) = U_{ij}((y_{ij} + z - a) \mod |A|), \qquad a \in A$$

such that $z \in A$ is sampled from U_{ij}. By using the mod operation we assume that $A = \{0, ..., |A| - 1\}$.

In addition to the proposed method, denoted as the soft-EM, we implemented a simplified decision method that ignores the reliability differences among the experts. The soft-majority procedure, denoted by 'soft-maj' is defined as follows:

$$\hat{x}_j = \arg\max_{a \in A} (\sum_{i=1}^{m} q_{ij}(a))$$

We also implemented algorithms based on a hard-decision process of the given soft-decision information. For each x_j let

$$\hat{y}_{ij} = \arg\max_{a \in A} q_{ij}(a)$$

be the most probable value of x_j according to expert i. We applied the EM procedure, described in Section 2 on this hard decision data and also a majority voting decision:

$$\hat{x}_j = \arg\max_{a \in A} (\sum_{i=1}^{m} 1_{\{\hat{y}_{ij} = a\}}).$$

We denote these two methods 'hard-EM' and 'hard-maj' respectively.

For each algorithm we measured the relative number of correct label prediction, i.e.:

$$\text{Score} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\hat{x}_j = x_j\}}. \tag{26}$$

We repeated the entire procedure described above 100 times. The average score as a function of the number of experts is shown in Fig. 1. Fig. 1 indicates that performance significantly improved by keeping information in a soft form and the best results were obtained by the proposed method.

To conclude, in this paper we developed an efficiently computed learning algorithm for integrating decisions from several unreliable experts. We showed the improved performance of the proposed approach compared to simplified methods that do not use the whole information. One possible future research direction would be to integrate the proposed algorithm with a classifier training algorithm (e.g. deep neural network) where the labeling is provided by a set of unreliable experts.

6. REFERENCES

[1] A. Alush and J. Goldberger. Ensemble segmentation using efficient integer linear programming. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 34:1966–1977, Oct. 2012.

[2] P. Donmez, J. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. *Knowledge Discovery and Data Mining (KDD)*, 2009.

[3] F. Parisi, F. Strino, B. Nadler, and Y. Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 114(4):1253–1258, 2014.

[4] V. C. Raykar, S. Yu, L. Zhao, A. Jerebko, C. Florin, G. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. *Intl. conf. on Machine Learning (ICML)*, 2009.

[5] 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, 2010.

[6] V.C. Raykar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13:491–518, 2012.

[7] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of Venus images. *Neural Information Processing Systems*, 1995.

[8] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imag.*, 23:903–921, 2004.

[9] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Neural Information Processing Systems*, 2010.

[10] J. 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 *Neural Information Processing Systems*, 2009.