Improving Learning-from-Crowds through Expert Validation

Mengchen Liu[1,4], **Liu Jiang**[1,4], **Junlin Liu**[1,4], **Xiting Wang**[2], **Jun Zhu**[3,4] **and Shixia Liu**[1,4*]
[1]School of Software, Tsinghua University, Beijing, P.R. China
[2]Microsoft Research, Beijing, P.R. China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, P.R. China
[4]Tsinghua National Lab for Information Science and Technology
{liumc13,jiangl16,liujl12,wang-xt11}@mails.tsinghua.edu.cn, {dcszj,shixia}@tsinghua.edu.cn

Abstract

Although several effective learning-from-crowd methods have been developed to infer correct labels from noisy crowdsourced labels, a method for post-processed expert validation is still needed. This paper introduces a semi-supervised learning algorithm that is capable of selecting the most informative instances and maximizing the influence of expert labels. Specifically, we have developed a complete uncertainty assessment to facilitate the selection of the most informative instances. The expert labels are then propagated to similar instances via regularized Bayesian inference. Experiments on both real-world and simulated datasets indicate that given a specific accuracy goal (e.g., 95%), our method reduces expert effort from 39% to 60% compared with the state-of-the-art method.

1 Introduction

Crowdsourcing has become one of the most cost-effective mechanisms to quickly obtain large amounts of labeled data [Wang and Zhou, 2016; Yan *et al.*, 2015; Zhou and He, 2016]. Such labeled data is the cornerstone of a variety of supervised learning methods [Zhang *et al.*, 2016]. However, due to individual differences among workers in terms of background, knowledge, and expertise, crowdsourced labels may be noisy and poor in quality. Researchers have developed a variety of effective learning-from-crowd algorithms to estimate correct labels from noisy data [Tian and Zhu, 2015; Zhou *et al.*, 2012]. Although these algorithms achieve some success in increasing accuracy, the unsupervised nature of these algorithms limits their performance.

Recently, Hung et al. [Hung *et al.*, 2015] introduced additional expert labels into these learning-from-crowd algorithms, and it is acknowledged as one of the pioneering efforts in this direction. For simplicity's sake, we denote this method as Hung's method. Although this method has demonstrated the effectiveness of leveraging expert labels at reducing the labor involved, it has two major issues: incomplete uncertainty assessment because it mainly considers the uncertainty caused by the data (e.g., input labels) and an indirect label propagation mechanism.

To address the above issues, we have developed a semi-supervised algorithm that simultaneously considers the uncertainty in each phase of machine learning and thoroughly propagate expert labels. A previous study has shown that uncertainty can be introduced in each phase of learning [Liu *et al.*, 2016; Wang and Zhai, 2016]. As a result, in the selection phase, we have developed a complete uncertainty assessment method for selecting the most informative instances, which jointly considers the uncertainty caused by the data, the model, and the solution, respectively. In the label propagation phase, the key is to seamlessly integrate the expert labels into a learning-from-crowd model to maximize their influence. To this end, we have formulated such integration as regularized Bayesian inference [Zhu *et al.*, 2014], along with a Gibbs sampler for performing Bayesian inference.

To demonstrate the method's effectiveness, we evaluated it with several simulated and real-world datasets. The experimental results show that for the best case, our method can achieve an accuracy of 95% with only 23% of the instances validated by an expert (Fig. 2E). Given a specific accuracy goal (e.g., 95%), our method reduces expert effort from 39% to 60% compared with Hung's method (Fig. 2A-D). In addition, the results clearly demonstrate that each of the two major components, the uncertainty assessment and the label propagation, can reduce the effort required.

The key technical contributions of this work include:

- **A complete uncertainty assessment method** that simultaneously considers the uncertainty in each phase of machine learning and in turn makes more gain.
- **An effective label propagation mechanism** that directly propagates the influence of expert labels via a formulation based on regularized Bayesian inference.

2 Background

Our label propagation algorithm is based on the max-margin majority voting (M^3V) model [Tian and Zhu, 2015], which is a state-of-the-art learning-from-crowd model to estimate the true labels from noisy data. This model introduces the concept of margin to improve its discriminative ability. The margin of an instance measures the separation between a potential correct label and any alternative label. This model maximizes

*S. Liu is the corresponding author.

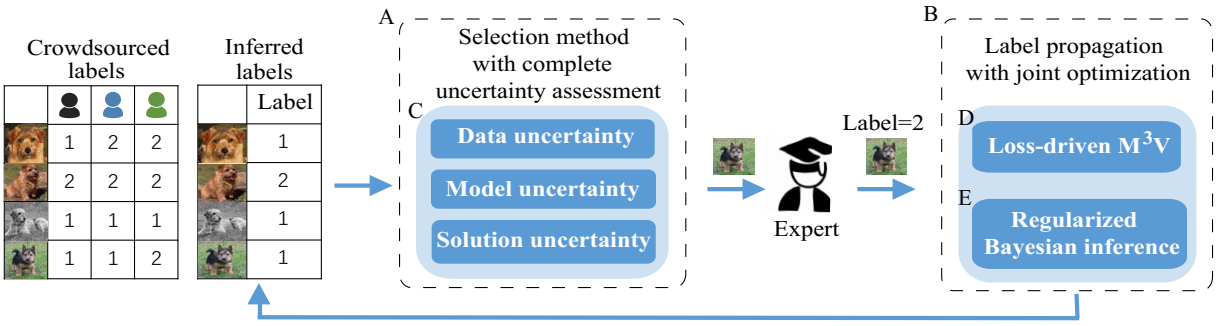

Figure 1: Basic idea of our method.

the overall margins of all instances and the posterior probability of correct labels $\mathbf{y}$ given the crowdsourced labels $\mathbf{X}^c$:

$$\inf_{q\in\mathcal{P}} \ \mathcal{L}(q(\boldsymbol{R})) \ + 2c\cdot\mathbb{E}_q\Big[\sum_{i=1}^{M}(\zeta_i)_+\Big], \tag{1}$$

where ζ_i measures the margin of instance i with a property that a larger ζ_i means a smaller margin. It can be computed by the method proposed by Crammer and Singer [Crammer and Singer, 2001]. $(x)_+ = \max(0, x)$, $\boldsymbol{R}$ is a set of random variables to be inferred (correct labels $\boldsymbol{y}$, worker reliability $\boldsymbol{\eta}$, and confusion tensor $\boldsymbol{\Phi}$), and $\mathcal{L}(q(\boldsymbol{R}))$ can be calculated by:

$$\mathcal{L}(q(\boldsymbol{R})) = KL(q(\boldsymbol{R})||p_0(\boldsymbol{R})) - \mathbb{E}_q\Big[\log(p(\boldsymbol{X^c}|\boldsymbol{R}))\Big]. \tag{2}$$

$p_0(\boldsymbol{R})$ is the prior. It is factorized as the product of a Dirichlet prior over $\boldsymbol{\Phi}$, a spherical Gaussian prior over $\boldsymbol{\eta}$, and a uniform prior over $\boldsymbol{y}$ [Tian and Zhu, 2015]. $p(\boldsymbol{X^c}|\boldsymbol{R})$ is the likelihood and is computed by the method proposed by Dawid and Skene [Dawid and Skene, 1979]. $p(\boldsymbol{R}|\boldsymbol{X^c})$ is the Bayesian posterior obtained by the standard Bayes' rule: $p(\boldsymbol{R}|\boldsymbol{X^c}) \propto p_0(\boldsymbol{R})p(\boldsymbol{X^c}|\boldsymbol{R})$. $q(\boldsymbol{R})$ is a general posterior. To distinguish it from $p(\boldsymbol{R}|\boldsymbol{X^c})$, $q(\boldsymbol{R})$ is called the desired post-data posterior. Its optimal distribution is obtained by solving the optimization problem defined in Eq. (1).

3 Method Overview

Given the crowdsourced labels $\mathbf{X}^c$ and expert labels $\mathbf{X}^e$, we model the inference of the correct labels $\mathbf{y}$ as a semi-supervised learning problem:

$$\mathbf{y} = f_A(\mathbf{X}^c, \mathbf{X}^e), \tag{3}$$

where the expert labels are the ground-truth labels for several selected instances, the crowdsourced labels are features of those instances, and $f_A(\cdot)$ is the aggregation function that combines the crowdsourced labels and the expert labels. In practice, the expert may make a mistake when labeling the selected instances. We follow Hung's lightweight confirmation check [Hung *et al.*, 2015] to make our algorithm robust with respect to the potential wrong expert labels.

In contrast to traditional semi-supervised learning, the ground-truth labels (expert labels) in our scenario are not ready and need to be acquired first. Because acquiring expert labels is expensive, our method iteratively selects the most informative instances to reduce the effort:

$$M_o^{(k)} = M_o^{(k-1)} \cup \underset{m\in M_s^{(k)}}{\operatorname{argmax}} \ I(m) \quad (k = 1, ..., n), \tag{4}$$

where $M_s^{(k)}$ is a subset that contains all the instances except those that are selected at iterations $1, ..., k-1$, and $I(\cdot)$ is an informativeness measure. The expert labels $\mathbf{X}^e$ are obtained by iteratively validating the instances in $M_o^{(k)}$.

Accordingly, we have developed a semi-supervised learning algorithm to incorporate expert labels. The input of our method are crowdsourced labels and the corresponding inferred correct labels. The inferred labels are computed by the M^{3}V model [Tian and Zhu, 2015]. To incorporate expert labels, we first use an instance selection method with a complete uncertainty assessment to find the most informative instances to be validated by an expert (Fig. 1A). Second, in the label propagation phase (Fig. 1B), the expert labels are propagated to other similar instances by using joint optimization. After the inferred correct labels are updated, more instances can be selected to further improve accuracy. The main features of the two steps are summarized as below.

- **Selection method.** We develop a complete uncertainty assessment that considers the uncertainty occurred in each phase of machine learning, i.e., data, model, and solution uncertainty (Fig. 1C). The selection strategy is then based on this complete uncertainty assessment, which makes more gain in quality control.
- **Label propagation.** By modeling the expert labels as a loss term in the M^{3}V model (loss-driven M^{3}V, Fig. 1D), we jointly optimize likelihood and other important factors such as margin during the propagation. We also show that this model can be solved by using regularized Bayesian inference (Fig. 1E).

4 Selection Method

The selection method aims to find the most informative instances to be validated by an expert. A widely-used framework to achieve this goal is uncertainty sampling [Settles, 2010]. This framework selects the instances that the learning algorithm is least certain about labeling. Accordingly, the key challenge is to accurately assess the uncertainty that can be introduced in each phase of learning [Wang and Zhai,

2016]. To address this challenge, we have developed a selection method that assesses the uncertainty associated with the data, the model, and the solution (inference process).

Data uncertainty. We follow Hung's method [Hung *et al.*, 2015] for computing data uncertainty. Specifically, two types of data uncertainty are considered: the uncertainty caused by noisy crowdsourced labels and the uncertainty caused by faulty workers. For each instance i, the uncertainty of its crowdsourced labels $U_c(i)$ is measured by using entropy-based information gain: $IG(i) = H(P) - H(P|i)$, where $H(P)$ is the sum of the Shannon entropies [Shannon, 2001] of all instances, and $H(P|i)$ measures the expected value of $H(P)$ if the label of instance i is known. The uncertainty introduced by faulty workers, $U_f(i)$, is defined as the total expected number of detected faulty workers if instance i is validated, which can be calculated by using Hung's worker-driven strategy.

Model uncertainty. To measure the uncertainty of the model, we compare the results of multiple learning-from-crowd models. For simplicity's sake, we use two learning-from-crowd models (A_1 and A_2) to illustrate the basic idea. For each instance i, we denote its label inferred by A_1 (A_2) as $y_i^{A_1}$ ($y_i^{A_2}$). A subset of instances is identified based on the disagreement of two models, which is defined as $M_d = \{i|y_i^{A_1} \neq y_i^{A_2}\}$. The model uncertainty of instance i is defined as $U_m(i) = \mathbf{1}_{M_d}(i)$, where $\mathbf{1}_{M_d}(i)$ is an indicator function. Next, we demonstrate that selecting instances with large model uncertainty makes more gain in quality control.

Theorem 1. *If the accuracy of A_1 is smaller than 1 and the accuracy of A_2 is larger than 0.5, then we have $p(i \in M_{s1}|i \in M_d) > p(i \in M_{s1}|i \in M)$, where M_{s1} represents the set of instances that are misclassified by A_1 and M represents the set that contains all the instances.*

Proof. We denote the accuracy of A_1, the accuracy of A_2, and $p(i \in M_{s1}, i \in M_{s2}, y_i^{A_1} \neq y_i^{A_2})$ as p_1, p_2 and p_3. With the assumption that A_1 and A_2 misclassify instances independently, we have $p(i \in M_{s1}|i \in M_d) = \frac{(1-p_1)p_2+p_3}{p_3+p_1(1-p_2)+p_2(1-p_1)}$ and $p(i \in M_{s1}|i \in M) = 1 - p_1$. Thus,

$$p(i \in M_{s1}|i \in M_d) > p(i \in M_{s1}|i \in M) \quad (5)$$

$$\Leftrightarrow \frac{(1-p_1)p_2 + p_3}{p_3 + p_1(1-p_2) + p_2(1-p_1)} > 1 - p_1 \quad (6)$$

$$\Leftrightarrow p_3 + (2p_2 - 1)(1 - p_1) > 0. \quad (7)$$

Inequation (7) can be easily proven by using the given conditions $p_1 < 1$ and $p_2 > 0.5$. □

The above theorem indicates that selecting instances from M_d instead of M increases the probability of selecting misclassified instances. Because correcting a misclassified instance usually brings more gain than verifying a correctly classified instance [Sun and Zhou, 2012], incorporating model uncertainty generally makes more gains. We can also employ more than two learning-from-crowd models based on the idea of query by disagreement [Cohn *et al.*, 1994]. Specifically, we first run all the models and estimate the performance of these models by using model likelihood. Then we regard the model that has the best performance as A_1 and the model that has the worst performance as A_2.

Solution uncertainty. Since we use Gibbs sampling to solve the model, we measure solution uncertainty by comparing the results of multiple Gibbs samplers. Specifically, the solution uncertainty $U_s(i)$ is calculated by using Shannon entropy: $-\sum_{l \in L} m_{i,l}/m \log(m_{i,l}/m)$. Here $m_{i,l}$ denotes the number of Gibbs samplers that return l as the estimated label for instance i and m ($m = 5$ in our implementation) is the number of Gibbs samplers used.

Combining different types of uncertainty. The aforementioned uncertainty types can be combined by a mixture model that samples instance i with probability $p(i) = \sum_{U_x \in \{U_c, U_f, U_m, U_s\}} p(i|U_x)p(U_x)$. Here $p(U_x)$ is the mixing coefficients and $p(i|U_x)$ denotes individual uncertainty component densities. We calculate the component density by using the exponential growth model [Wikipedia, 2017], $p(i|U_x) \propto e^{U_x(i)}$, which aims to increase the sampling probability of instances with larger uncertainty values. In our implementation, $p(U_c)$, $p(U_f)$, $p(U_m)$, and $p(U_s)$ are set to 0.36, 0.04, 0.54, and 0.06, respectively.

5 Label Propagation

Label propagation aims to maximize the influence of expert labels by propagating them to other unconfirmed instances. The key challenge is to effectively propagate expert labels while decreasing the effect caused by the noisy crowdsourced labels. The state-of-the-art method [Hung *et al.*, 2015] used an indirect propagation mechanism that may fail to effectively propagate expert labels. In their method, the expert labels are first used to assess the reliability of workers, and then the reliability is exploited to infer the correct labels of other instances that are labeled by these workers. Because each worker only labels a small set of instances, the influence of expert labels is limited in such an indirect propagation. In addition, this indirect propagation process may accumulate more errors due to the noise in the crowdsourced labels. Thus, it is desirable to develop a more direct way to maximally and accurately propagate expert labels to other instances.

An intuitive method to directly propagate expert labels is using the coin-toss model [Goldberg, 1989]. In this model, the label of a validated instance is propagated to similar instances with a probability p. Specifically, for each instance i, its most similar validated instance i_v is selected. If i is not validated, we set its label the same as that of i_v with probability $p = s_{i_v,i}$, where $s_{i_v,i}$ is the similarity between i and i_v. However, this method only considers the similarity between two instances without taking the confidence of the inferred labels into account. As a result, the coin-toss model may lead to unintentional modifications of correctly classified instances with high confidence.

To solve this problem, we have developed a loss-driven algorithm based on the M^{3}V model [Tian and Zhu, 2015]. The main feature of our algorithm is that it jointly considers the influence of expert labels and other important factors such as the likelihood of crowdsourced labels.

Expert labels as a loss term. Based on the M^{3}V model, we incorporate expert labels into the original optimization

function of M^{3}V as a loss term. The loss term aims to maximally propagate the influence of expert labels with the following two constraints. First, for validated instances, the loss term ensures their labels are correctly set. Second, for other instances, the loss term penalizes the model for setting its label deviating far from its similar validated instances. Accordingly, we add a loss term composed of two parts (L_1 and L_2), one for each goal:

$$\inf_{q\in\mathcal{P}} \ \mathcal{L}(q(\boldsymbol{R})) + 2c\cdot\mathbb{E}_q\Big[\sum_{i=1}^{M}(\zeta_i)_+\Big] + \alpha\cdot L_1 + \beta\cdot L_2, \quad (8)$$

$$L_1 = \mathbb{E}_q\Big[\sum_{i\in\mathbb{S}}\|y_i - l_i\|_0\Big], \quad (9)$$

$$L_2 = \mathbb{E}_q\Big[\sum_{i\notin\mathbb{S}} s_{i,i_v}\|y_i - l_{i_v}\|_0\Big], \quad (10)$$

where $\mathbb{S}$ is the set of validated instances, l_i is the expert label of instance i, $s_{i,j}$ is the similarity between i and j, and α, β are regularization factors that control the influence of L_1 and L_2, respectively.

As shown in Eq. (8), our model jointly optimizes the influence of expert labels, the likelihood of a crowdsourced label, and the margin of an instance.

Regularized Bayesian inference. Solving this optimization problem can be viewed as performing a regularized Bayesian inference. To perform this inference, we develop a Gibbs sampler by absorbing the loss term into $\mathcal{L}(q(\boldsymbol{R}))$. Specifically, we first obtain the following equation by substituting Eqs. (9) and (10) into Eq. (8):

$$\begin{aligned}\inf_q\ & KL(q(\boldsymbol{R})||p_0(\boldsymbol{R})) - \int q(\boldsymbol{R})\ \log p(\boldsymbol{X^c}|\boldsymbol{R})d\boldsymbol{R}\\ & - \int q(\boldsymbol{R})\ \log\exp(-\sum_{i=1}^{M} 2c(\zeta_i)_+)d\boldsymbol{R}\\ & - \int q(\boldsymbol{R})\ \log\exp(-\sum_{i\in\mathbb{S}}\alpha\ \|y_i - l_i\|_0)d\boldsymbol{R}\\ & - \int q(\boldsymbol{R})\log\exp(-\sum_{i\notin\mathbb{S}}\beta\ s_{i,i_v}\|y_i - l_{i_v}\|_0)d\boldsymbol{R},\end{aligned} \quad (11)$$

by grouping the last four terms, Eq. (11) is rewritten as:

$$\inf_q\ KL(q(\boldsymbol{R})||p_0(\boldsymbol{R})) - \mathbb{E}_q\Big[\log(\tilde{p}(\boldsymbol{X^c},\boldsymbol{R}))\Big], \quad (12)$$

where

$$\begin{aligned}&\tilde{p}(\boldsymbol{X^c},\boldsymbol{R}) = p(\boldsymbol{X^c}|\boldsymbol{R})\psi(\boldsymbol{y}|\boldsymbol{X^c},\boldsymbol{\eta})\sigma(\boldsymbol{y})\tau(\boldsymbol{y}),\\ &\psi(\boldsymbol{y}|\boldsymbol{X^c},\boldsymbol{\eta}) = \exp(-\sum_{i=1}^{M} 2c(\zeta_i)_+),\\ &\sigma(\boldsymbol{y}) = \exp(-\sum_{i\in\mathbb{S}}\alpha\ \|y_i - l_i\|_0),\\ &\tau(\boldsymbol{y}) = \exp(-\sum_{i\notin\mathbb{S}}\beta\ s_{i,i_v}\|y_i - l_{i_v}\|_0).\end{aligned} \quad (13)$$

Here, $\tilde{p}(\boldsymbol{X^c},\boldsymbol{R})$ is an unnormalized pseudo-likelihood. According to the theory of regularized Bayesian inference [Zhu *et al.*, 2014], solving Eq. (12) is equivalent to sampling from:

$$\begin{aligned}\hat{q}(\boldsymbol{R}) &= p_0(\boldsymbol{R})\tilde{p}(\boldsymbol{X^c},\boldsymbol{R})\\ &= p_0(\boldsymbol{R})p(\boldsymbol{X^c}|\boldsymbol{R})\psi(\boldsymbol{y}|\boldsymbol{X^c},\boldsymbol{\eta})\sigma(\boldsymbol{y})\tau(\boldsymbol{y}).\end{aligned} \quad (14)$$

By introducing an augmented variable $\boldsymbol{\lambda}$ and exploiting the property that the correct label of each instance is independent from each other, we obtain:

$$\hat{q}(\boldsymbol{R}) \propto p_0(\boldsymbol{R})\prod_{i=1}^{M} p(\boldsymbol{x}_i|\boldsymbol{\Phi},y_i)\psi(y_i|\boldsymbol{x}_i,\boldsymbol{\eta})\sigma(y_i)\tau(y_i), \quad (15)$$

where

$$\psi(y_i|\boldsymbol{x_i},\boldsymbol{\eta}) = \int_0^{\infty}\psi(y_i,\lambda_i|\boldsymbol{x_i},\boldsymbol{\eta})d\lambda_i, \quad (16)$$

$$\psi(y_i,\lambda_i|\boldsymbol{x_i},\boldsymbol{\eta}) = (2\pi\lambda_i)^{-\frac{1}{2}}\exp(\frac{-1}{2\lambda_i}(\lambda_i + c\zeta_i)^2). \quad (17)$$

By using the Bayes' rule, the new conditional distribution to sample the correct labels $\boldsymbol{y}$ can be computed by:

$$q(y_i|\boldsymbol{\Phi},\boldsymbol{\eta},\lambda_i) \propto p(\boldsymbol{x_i}|\boldsymbol{\Phi},y_i)\psi(y_i,\lambda_i|\boldsymbol{x_i},\boldsymbol{\eta})\sigma(y_i)\tau(y_i). \quad (18)$$

To compute the conditional distribution of the remaining variables ($\boldsymbol{\Phi},\boldsymbol{\eta},\boldsymbol{\lambda}$), we follow the method proposed in the M^{3}V model. The time complexity of our method is the same as that in M^{3}V because we do not sample additional variables.

6 Evaluation

We conducted three experiments to demonstrate the effectiveness of our method. The first one briefly evaluates the overall performance. The results show that our method significantly reduces expert effort compared with Hung's method. The second and third experiments analyze the selection method and the label prorogation method. All the experiments were conducted on a workstation with Intel Core i5 CPU (3.3 GHz) and 16 GB of Memory.

6.1 Experimental Settings

Datasets. We used the following datasets in our experiments.

- **Dog** [Zhou *et al.*, 2012]: It contains 800 images of 4 breeds of dogs from ImageNet [Deng *et al.*, 2009]. Each image is labeled by 10 workers.
- **Age** [Han *et al.*, 2015]: It contains 1,002 face images. The workers were asked to estimate the age of the person in each image, which is discretized into 7 bins.
- **Monkey**: It contains images of 4 kinds of wild monkeys (Siamang, Guenon, Patas and Baboon). These images were selected from ImageNet and we simulated the crowdsourced labels for each image by the method used in [Hung *et al.*, 2013].
- **News**: It contains documents of 4 topics from the 20NewsGroup dataset [Lang, 1995]. We simulated the crowdsourced labels for each document by the method used in [Hung *et al.*, 2013].

Datasets		Instances	Workers	Classes
Real-world	Dog	800	109	4
	Age	1002	165	7
Simulated	Monkey	957	104	4
	News	2007	186	4

Table 1: Datasets statistics. In the simulated datasets, the images and documents are real, but the crowdsourced labels are simulated.

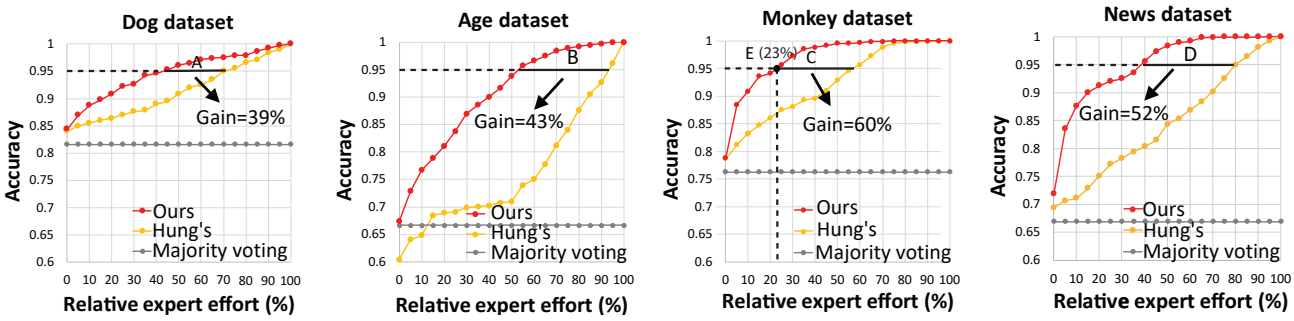

Figure 2: Comparison of accuracy and relative expert effort on multiple datasets (Gain = (Hung's - Ours)/Hung's).

Statistics of all the datasets are summarized in Table 1.

Expert labels. The ground-truth labels in each dataset are treated as the expert labels in our experiments.

Feature extraction. To measure the similarity between instances, we extracted a feature vector for each instance. For an image, we extracted its feature vector by using a deep convolutional neural network: VGG-NET [Simonyan and Zisserman, 2014]. We used the output of the last but one fully-connected layer as the feature vector of the image. For a document, we extracted its feature vector by using TF-IDF.

Criteria. We compared our method with the baseline methods in terms of the following criteria.

- **Relative expert effort** is defined as N_s/N, where N_s represents the number of validated instances and N is the total number of instances in the dataset.
- **Accuracy** is defined as N_c/N, where N_c is the number of correctly classified instances.

6.2 Overall Performance

To evaluate the overall performance of our method, we first compared it with Hung's method in terms of accuracy and expert effort. Fig. 2 shows how accuracy changes with relative expert effort for both methods on four datasets. Note that the base model in Hung's method, the DS (Dawid-Skene) model [Dawid and Skene, 1979], is different from ours, the M^{3}V model. Therefore, when expert effort is 0, the performance of the two methods is different. Our method performs better than theirs on most of the four datasets. By analyzing the results, we have the following conclusions.

First, compared with Hung's method, our method is able to achieve much higher accuracy at most levels of expert effort on all datasets. This demonstrates that our method can significantly reduce expert effort. In particular, given a specific accuracy goal, 95%, our method reduces expert effort from 39% to 60% compared with Hung's method (Fig. 2A-D). Among the four datasets, our method performs the best on the Age dataset. This is because this dataset contains some outliers (e.g., the young person looks very old). The M^{3}V model adopted by our method avoids the over-fitting problem by adding margin as a regularization term.

Second, as expert effort increases, the accuracy of our method first increases very fast and then slows down as relative expert effort approaches 100%. This indicates that our method can select the most informative instances in the beginning and correctly propagate the labels to other instances. As a result, the accuracy of our method can be considerably improved even when the number of expert labels is small. This is a very desirable feature for many real-world applications because the expert labels are usually expensive to acquire.

We then analyzed how the estimated assignment probability of the correct labels changed with different levels of expert effort (0%, 15%, and 30%). For each instance, a good method assigns a higher probability to the correct label than to an incorrect one. Fig. 3 shows two histograms of the probability distribution of the News dataset and the Dog dataset. For each instance, we examined the assignment probability $\hat{q}(\boldsymbol{R})$ (Sec. 5) of its correct label. If this assignment probability is in a probability bin, the count for that bin is increased by 1. We noted that more than 10% instances had a very low probability (<0.1) to be correctly classified when no expert labels were incorporated. By incorporating more expert labels, fewer instances have a very low probability to be correctly classified. We also observed that the number of instances that had a very high probability (>0.9) to be correctly classified increased as the expert labels increased. This result demonstrates that the performance of the learning-from-crowd method is improved after more expert labels are provided.

6.3 Selection Method

Baselines. Four baselines are used in this experiment. The first baseline (**Hung's selection method**) is the selection method used in Hung's method, which takes into account data uncertainty. The second baseline (**Model uncertainty**) only considers model uncertainty. The third baseline (**Solution uncertainty**) only considers solution uncertainty. The last baseline (**Random**) randomly selects instances from M (Sec. 4). To eliminate the bias caused by different label propaga-

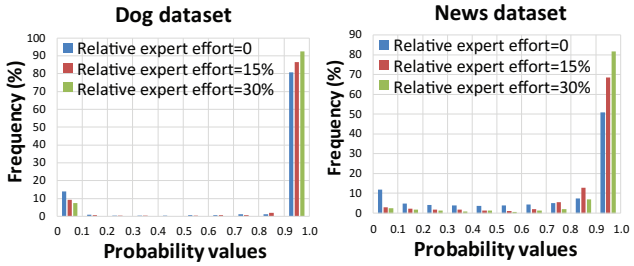

Figure 3: Probability distribution of correct labels

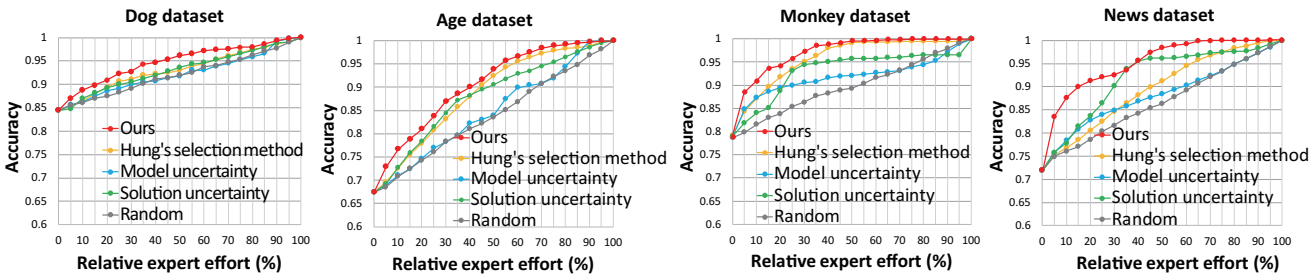

Figure 4: Comparison of different selection methods. For a fair comparison, our label propagation method and the M^{3}V model are adopted in all the methods.

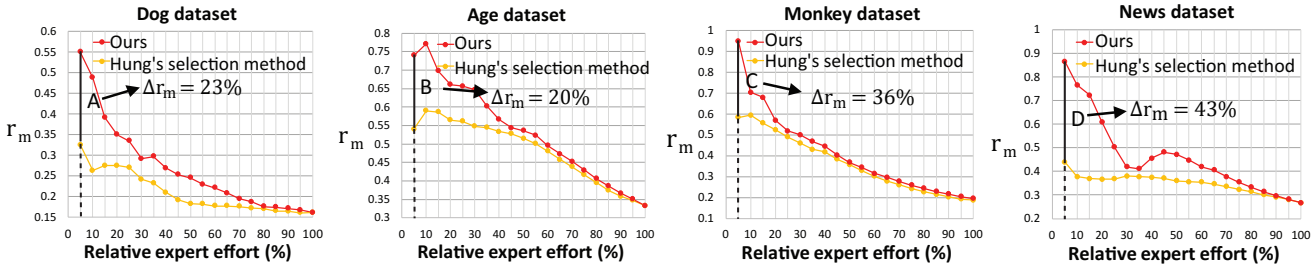

Figure 5: Comparison of the misclassified ratio (the number of the misclassified instances over that of the selected instances).

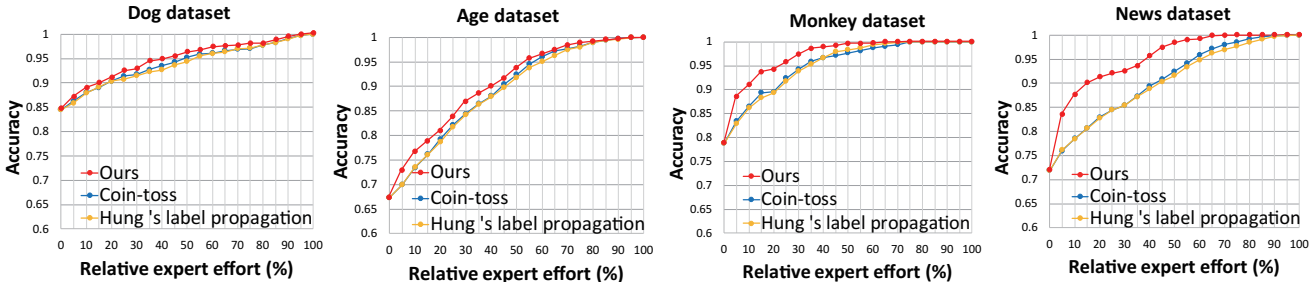

Figure 6: Comparison of different label propagation algorithms. For a fair comparison, all the algorithms employ the same base model, M^{3}V, and our selection method.

tion methods, we combined these selection methods with our label propagation method and then evaluated the corresponding accuracy and expert effort. In all the experiments, we provided multiple instances to the expert each time.

Results. As shown in Fig. 4, our selection method achieves higher accuracy compared with the baselines on all datasets. This result demonstrates that integrating model uncertainty and solution uncertainty makes more gain.

Next, we demonstrated that our selection method could select more misclassified instances than Hung's method for different levels of expert effort. To this end, we defined the misclassified ratio, r_m, as the number of selected misclassified instances over that of total selected instances. As shown in Fig. 5, our selection method increases the ratio from 20% to 43% at a low level of expert effort (5%) (Fig. 5A-D). The reason is that in the beginning we select the instances among M_d (Sec. 4) and thus our method can select more misclassified instances. As expert effort increases, the improvement gradually decreases because all the instances in M_d have been selected and the method starts to select the instances from $M \setminus M_d$. This result clearly illustrates the effectiveness of incorporating both model uncertainty and solution uncertainty.

6.4 Label Propagation

Baselines. The adopted baselines are Hung's method [Hung *et al.*, 2015] and the coin-toss model. To fairly compare our label propagation method with Hung's method, we adopted the same base model (M^{3}V) and our selection method.

Results. As shown in Fig. 6, our method outperforms Hung's method and the coin-toss model. Compared with the indirect propagation of Hung's method, our method is able to directly propagate expert labels to more similar instances. In this way, our method is less affected by the noise in the crowdsourced labels and thus achieves higher accuracy. The results also show that the coin-toss model sometimes leads to unintentional modifications by only considering the similarity between instances. For example, in the Monkey dataset, the accuracy of the coin-toss model remains the same when relative expert effort is 15% and 20%. In contrast, our method achieves higher accuracy as expert effort increases.

7 Related Work

Most recent learning-from-crowd methods can be categorized into three groups: generative methods, discriminative methods, and hybrid methods.

Generative methods use a probabilistic model to generate the crowdsourced labels conditioned on the correct labels and assumptions of worker behavior. A typical example is the DS (Dawid-Skene) model [Dawid and Skene, 1979], in which the behavior of each worker is modeled by a confusion matrix. To improve the performance of the DS model, researchers later put a prior over confusion matrices [Liu *et al.*, 2012] and considered task difficulties [Bachrach *et al.*, 2012].

Discriminative methods directly resolve the correct labels via some aggregation rules. An intuitive example is majority voting [Snow *et al.*, 2008]. Another example is weighted majority voting [Karger *et al.*, 2011], which distinguishes a spammer from a reliable worker (worker reliability).

Recently, Tian et al. [Tian and Zhu, 2015] developed a hybrid method to combine the discriminative ability of discriminative methods and the flexibility of generative methods. In particular, they extended the weighted majority voting model with the notion of margin, and then coupled it with the DS model via regularized Bayesian inference.

Although these algorithms have increased the accuracy, the performance is still limited by the unsupervised nature of these algorithms. To solve this problem, Hung et al. [Hung *et al.*, 2015] proposed a method to minimize expert workload for validating crowdsourced labels. However, this method suffers from two major issues: incomplete uncertainty assessment and an indirect label propagation mechanism. Compared to their work, our selection method considers the uncertainty in each phase of machine learning, which is able to select more misclassified instances and makes more gain. In addition, we directly propagate the influence of expert labels by formulating the label propagation problem as regularized Bayesian inference.

8 Conclusions and Future Work

In this paper, we have presented a semi-supervised algorithm to improve the quality of crowdsourced labels by incorporating expert labels. Our method comprehensively considers the uncertainty occurred in each phase of machine learning to select the most informative instances. Moreover, it models the influence of expert labels to other instances with regularized Bayesian inference. Experimental results demonstrate that our method reduces the effort of experts from 39% to 60% compared with the state-of-the-art method.

Directions for future investigation include the reliability evaluation of expert labels. Currently, we assume that the expert labels are the ground-truth labels. However, the expert may make a mistake due to fatigue or lack of knowledge. For cases where more than one experts validate the estimated labels, the experts may not always make agreement on each label. Accordingly, it would be interesting to investigate the influence of expert disagreement on the final result. Another important topic for future research is to investigate the possibility of incorporating other learning-from-crowd algorithms [Zhou *et al.*, 2012; 2014]. This will facilitate the improvement of model uncertainty and form an ensemble of estimators for boosting accuracy improvement.

Acknowledgments

M. Liu, L. Jiang, J. Liu, and S. Liu are supported by National NSF of China (No. 61672308). J. Z is supported by NSFC Projects (Nos. 61620106010, 61621136008) and the Youth Top-notch Talent Support Program. The authors would like to thank Tian Tian for providing the code of the M^3V model.

References

[Bachrach *et al.*, 2012] Yoram Bachrach, Thore Graepel, Tom Minka, and John Guiver. How to grade a test without knowing the answers — a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *ICML*, pages 1183–1190, 2012.

[Cohn *et al.*, 1994] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[Crammer and Singer, 2001] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2(Dec):265–292, 2001.

[Dawid and Skene, 1979] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, 28(1):20–28, 1979.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Goldberg, 1989] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[Han *et al.*, 2015] Hu Han, Charles Otto, Xiaoming Liu, and Anil K Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE PAMI*, 37(6):1148–1161, 2015.

[Hung *et al.*, 2013] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. An evaluation of aggregation techniques in crowdsourcing. In *ICWISE*, pages 1–15, 2013.

[Hung *et al.*, 2015] Nguyen Quoc Viet Hung, Duong Chi Thang, Matthias Weidlich, and Karl Aberer. Minimizing efforts in validating crowd answers. In *SIGMOD*, pages 999–1014, 2015.

[Karger *et al.*, 2011] David R. Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *NIPS*, pages 1953–1961, 2011.

[Lang, 1995] Ken Lang. Newsweeder: Learning to filter netnews. In *ICML*, pages 331–339, 1995.

[Liu *et al.*, 2012] Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. In *NIPS*, pages 692–700. 2012.

[Liu *et al.*, 2016] Mengchen Liu, Shixia Liu, Xizhou Zhu, Qinying Liao, Furu Wei, and Shimei Pan. An uncertainty-aware approach for exploratory microblog retrieval. *IEEE TVCG*, 22(1):250–259, 2016.

[Settles, 2010] Burr Settles. Active learning literature survey. 52(55-66):11, 2010.

[Shannon, 2001] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE MC2R*, 5(1):3–55, 2001.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[Snow *et al.*, 2008] 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*, pages 254–263, 2008.

[Sun and Zhou, 2012] Tengyu Sun and Jie Zhou. Batch-mode active learning by using misclassified data. In *ALRA Workshop at ECML-PKDD*, 2012.

[Tian and Zhu, 2015] Tian Tian and Jun Zhu. Max-margin majority voting for learning from crowds. In *NIPS*, pages 1621–1629, 2015.

[Wang and Zhai, 2016] Xizhao Wang and Junhai Zhai. *Learning with Uncertainty*. CRC Press, 2016.

[Wang and Zhou, 2016] Lu Wang and Zhi-Hua Zhou. Cost-saving effect of crowdsourcing learning. In *IJCAI*, pages 2111–2117, 2016.

[Wikipedia, 2017] Exponential growth model. `https://en.wikipedia.org/wiki/Exponential_growth`, February 2017.

[Yan *et al.*, 2015] Rui Yan, Yiping Song, Cheng-Te Li, Ming Zhang, and Xiaohua Hu. Opportunities or risks to reduce labor in crowdsourcing translation? characterizing cost versus quality via a pagerank-hits hybrid model. In *IJCAI*, pages 1025–1032, 2015.

[Zhang *et al.*, 2016] Jing Zhang, Xindong Wu, and Victor S Sheng. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46(4):1–34, 2016.

[Zhou and He, 2016] Yao Zhou and Jingrui He. Crowdsourcing via tensor augmentation and completion. In *IJCAI*, pages 2435–2441, 2016.

[Zhou *et al.*, 2012] Denny Zhou, Sumit Basu, Yi Mao, and John C Platt. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, pages 2195–2203, 2012.

[Zhou *et al.*, 2014] Dengyong Zhou, Qiang Liu, John C Platt, and Christopher Meek. Aggregating ordinal labels from crowds by minimax conditional entropy. In *ICML*, pages 262–270, 2014.

[Zhu *et al.*, 2014] Jun Zhu, Ning Chen, and Eric P. Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *JMLR*, 15(1):1799–1847, 2014.