Crowdsourcing with Meta-Knowledge Transfer (Student Abstract)

Sunyue Xu, Jing Zhang*

School of Computer Science and Engineering, Nanjing University of Science and Technology
200 Xiaolingwei Street, Nanjing 210094, China
{sue, jzhang}@njust.edu.cn

Abstract

When crowdsourced workers perform annotation tasks in an unfamiliar domain, their accuracy will dramatically decline due to the lack of expertise. Transferring knowledge from relevant domains can form a better representation for instances, which benefits the estimation of workers' expertise in truth inference models. However, existing knowledge transfer processes for crowdsourcing require a considerable number of well-collected instances in source domains. This paper proposes a novel truth inference model for crowdsourcing, where (meta-)knowledge is transferred by meta-learning and used in the estimation of workers' expertise. Our preliminary experiments demonstrate that the meta-knowledge transfer significantly reduces instances in source domains and increases the accuracy of truth inference.

Introduction

Crowdsourcing provides a fast and low-cost solution to collect annotations for data. To address the imperfectness of crowdsourced labels, researchers in the AI community adopt a repeated labeling and inference scheme, where an instance is labeled by different workers and the potential true labels are estimated by statistical inference models (Zhang, Wu, and Sheng 2016). Although truth inference models have improved the quality of labels, their performance could be guaranteed only under the condition that most crowdsourced labels should be correct (Awasthi et al. 2017). However, this condition does not always hold in reality, facing completely new tasks, the majority of workers may do not have adequate expertise to make the right judgments so that massive noises will appear in crowdsourced labels, resulting in the malfunction of inference algorithms.

Transfer learning is thought to have a bright promise to address the above issue. Knowledge obtained from relevant domains may provide beneficial hints in truth inference without the awareness of workers. The earliest attempt of this idea was in (Fang, Yin, and Tao 2014), where a probabilistic inference model was proposed to transfer knowledge from auxiliary domains to generate a new representation for instances. The new representation of instances will make the estimation of workers' expertise more accurate so that the performance of the truth inference model will be improved. Han *et al.* (2020) followed the same train of thought to learn a better representation of instances from the knowledge transferred from multiple source domains, which improves the inference performance in the multi-class scenario. However, these methods also have a weakness. To obtain a good knowledge transfer effect, we should in advance prepare abundant source-domain instances, whose features also should be similar to those in the target domain. The preparation of the source-domain instances will cost much and occasionally be infeasible.

This paper proposes a novel crowdsourcing model with meta-knowledge transfer (CrowdMeta). CrowdMeta utilizes meta-learning to transfer knowledge from source domains. The transferred knowledge that forms a new representation for instances is called *meta-knowledge*, which summarizes the high-level knowledge of how to learn. With the help of a model pre-trained from a small number of instances, we only need to fine-tune the model with a few instances, then we can obtain a good knowledge transfer effect. The higher-level meta-knowledge (as the representation of instances) is used to model the workers' expertise better in truth inference. Our preliminary experimental results on real-world datasets show that CrowdMeta outperforms existing knowledge-transfer crowdsourcing models.

The Proposed Method

Transfer Learning Process

The purpose of transfer learning is to find an internal representation suitable for tasks (instances), which helps model the professional ability of crowdsourced workers. We adopt the idea of few-shot meta-learning to pre-train a model, which only needs a small number of instances and trials to quickly adapt to new tasks (Finn, Abbeel, and Levine 2017).

In our meta-learning settings, the training instances $\mathcal{T}$ obey the distribution $p(\mathcal{T})$. Function f_θ is a model presented by parameters θ. During meta-training, the model parameters θ are updated by multiple rounds of gradient descent optimization. In the t-th round of optimization, we have

$$\theta_t^{'} = \theta - \alpha \bigtriangledown_\theta \mathcal{L}_{\mathcal{T}_t}(f_\theta), \tag{1}$$

where $\mathcal{L}$ denotes a loss function and the step size α can be fixed as a hyperparameter. We optimize the performance of

*The corresponding author is Dr. Jing Zhang.

model $f_{\theta_t'}$ by the updated parameters $\theta_t^{'}$ through $\mathcal{T}_t$. The optimization objective is as follows:

$$\min_{\theta} \sum_{\mathcal{T}_t} \mathcal{L}_{\mathcal{T}_t}(f_{\theta_t'}) = \sum_{\mathcal{T}_t} \mathcal{L}_{\mathcal{T}_t}(f_{\theta-\alpha\nabla_\theta \mathcal{L}_{\mathcal{T}_t}(f_\theta)}). \quad (2)$$

And then we used K instances in the target domain to fine-tune the pre-trained model. The K instances used in this step have similar class and feature distributions to the new target instances. Finally, we use the fine-tuned model to extract the high-level representation $\mathbf{V}$ for the target instances, which is thought to be more essential and critical in truth inference.

Crowdsourcing Process

Given a set of N annotation tasks (instances), W crowdsourced workers provide noisy labels for the tasks, denoted by $\{l_{w,i}\}_{w=1,i=1}^{W,N}$. We use a variable $\mathbf{e}^w$ to denote the expertise of worker w, and let it associate with instance representation $\{\mathbf{v}_i\}_{i=1}^{N}$. We define the label quality q_i^w of worker w on task i as $q_i^w = \mathbf{v}_i \cdot \mathbf{e}^w$. The conditional probability of a worker's expertise on an instance representation (i.e., the normalized label quality) is defined by logistic regression as

$$p(q_i^w \mid \mathbf{v}_i) = (1 + \exp(-q_i^w))^{-1}. \quad (3)$$

The joint probability of hidden true labels of all the instances can be calculated as follows:

$$p = \prod_{i}^{N} p(y_i \mid \mathbf{v}_i) \prod_{w}^{W} p(q_i^w \mid \mathbf{v}_i) p(l_{w,i} \mid y_i, q_i^w). \quad (4)$$

The observed variables include the crowdsourced labels $\{l_{w,i}\}_{w=1,i=1}^{W,N}$ and the high-level representation of the target instances. The crowdsourcing learning objective is to infer a group of hidden variables $\Phi = \{\{y_i\}_{i=1}^{N}, \{\mathbf{e}^w\}_{w=1}^{W}\}$, which can be solved by an EM algorithm.

E-step We compute the posterior on the estimated true labels using the current parameters:

$$\begin{aligned}\hat{p}(y_i) &= p(y_i \mid \mathbf{e}_i, \mathbf{l}_i, \mathbf{v}_i) \propto p(y_i, \mathbf{e}_i, \mathbf{l}_i \mid \mathbf{v}_i) \\ &= p(y_i \mid \mathbf{v}_i) \prod_{w}^{W} p(l_{w,i} \mid y_i, q_i^w) p(q_i^w \mid \mathbf{v}_i).\end{aligned} \quad (5)$$

M-step We update the model parameters by maximizing the posterior expectation of the variables:

$$\begin{aligned}\Phi &= \max \mathbb{E}_y \left[\log(p(\mathbf{e}_i, \mathbf{l}_i, \mathbf{v}_i \mid y_i))\right] \\ &= \max \sum_{w}^{W} \sum_{i}^{N} \mathbb{E}_{y_i} \left[\log(q_i^w \mid \mathbf{v}_i)\right] \\ &+ \log p(y_i \mid \mathbf{v}_i) + \log p(l_{w,i} \mid y_i, \mathbf{e}_i).\end{aligned} \quad (6)$$

We can use the L-BFGS algorithm to optimize the above auxiliary function and update the parameters in a standard EM iterative procedure.

Experiments

We compared the proposed CrowdMeta with MV, CrowdTrU (Fang, Yin, and Tao 2014), and CrowdMKT (Han et al. 2020) on simulated crowdsourcing datasets. Our simulation was performed on the MiniImagenet dataset. We randomly selected 5×100 instances from the dataset as the target tasks. We simulated four types of workers (spammer, random, normal, and expert), using 30 workers to provide labels for the tasks, with an average of three labels for each.

	Experiment 1	Experiment 2
Majority Vote (MV)	0.676	0.696
CrowdTrU	0.684	0.712
CrowdMKT	0.704	0.720
CrowdMeta (Ours)	**0.728**	**0.732**

Table 1: Comparison of the performance of truth inference (in terms of accuracy) under two experimental settings.

The two experiments randomly selected different instances as the target tasks. In the first one, the source domain of CrowdTrU and CrowdMKT have the same number of instances as CrowdMeta does, with 5 instances per class. In the second one, the source domain of CrowdTrU and CrowdMKT used more instances, i.e., 60 instances per class of each source domain recommended by CrowdMKT. CrowdMeta still keeps 5 instances per class.

Table 1 shows the simulation results of truth inference in terms of accuracy. Overall, the models using transfer learning outperform the baseline MV. Our CrowdMKT performs the best because it extracts more inherent features of the tasks, which is more effective than directly amplifying the shared features of the source and target domains. That is, CrowdTrU and CrowdMKT are likely to amplify the noises at the same time. CrowdMTK can benefit from the increase of the instances in source domains, but its performance is still difficult to exceed that of our CrowdMeta.

Future Work

We will further optimize the crowdsourcing models and the meta-knowledge transfer methods as well as introduce active learning to allocate tasks and workers more efficiently.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2018AAA0102002 and the National Natural Science Foundation of China under Grants 62076130 and 91846104.

References

Awasthi, P.; Blum, A.; Haghtalab, N.; and Mansour, Y. 2017. Efficient PAC learning from the crowd. In *COLT*, 1–24.

Fang, M.; Yin, J.; and Tao, D. 2014. Active learning for crowdsourcing using knowledge transfer. In *AAAI*, 1809–1815.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 1126–1135.

Han, G.; Tu, J.; Yu, G.; Wang, J.; and Domeniconi, C. 2020. Crowdsourcing with multiple-source knowledge transfer. In *IJCAI*, 2908–2914.

Zhang, J.; Wu, X.; and Sheng, V. S. 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46(4): 543–576.