Who knows best? A Case Study on Intelligent Crowdworker Selection via Deep Learning

Marek Herde*, Denis Huseljic, Bernhard Sick, Ulrich Bretschneider and Sarah Oeste-Reiß

University of Kassel, Wilhelmshöher Allee 73, Kassel, 34121, Germany

Abstract

Crowdworking is a popular approach for annotating large amounts of data to train deep neural networks. However, parts of the annotations are often erroneous. In a case study, we demonstrate how an intelligent crowdworker selection via deep learning reduces the number of erroneous annotations and, thus, the annotation costs of obtaining reliable data for training deep neural networks.

1. Introduction

Deep neural networks (DNNs) typically need large amounts of annotated data to make reliable predictions in supervised learning tasks [1]. Crowdworking collects annotations by requesting crowdworkers to solve microtasks [2], such as image classifications. The crowdworkers mostly receive payments as compensation, leading to high costs for massive datasets. Parts of the annotations may be erroneous because crowdworkers are error-prone for various causes [3, 4], e.g., missing knowledge. Thus, many crowd-learning techniques have been proposed to train well-performing DNNs despite annotations from error-prone crowdworkers [5, 6, 7, 8]. They abstract from the specific error causes to jointly estimate crowdworkers' performances and instances' true annotations. Commonly, these techniques are employed after completing a crowdworking campaign. However, leveraging the crowdworkers' performance estimates to optimize an ongoing campaign appears beneficial. Therefore, this article studies whether crowd-learning techniques can answer the question *"Who knows best?"* to select crowdworkers intelligently. In a case study with classification data, we show that such a crowdworker selection reduces the number of erroneous annotations and allows us to train DNNs with lower misclassification rates than a random selection of crowdworkers at the same annotation costs.

This article targets a subfield of machine learning to support crowdworking [9], including active learning [10]. Compared to active learning for crowdworking [11, 12], we focus on studying the potential of state-of-the-art crowd-learning techniques to improve crowdworker selection and outline challenges when employing such techniques in real crowdworking campaigns.

IAL@ECML-PKDD'23: 7th Intl. Worksh. & Tutorial on Interactive Adaptive Learning, Sep. 22nd, 2023, Torino, Italy
*Corresponding author.
marek.herde@uni-kassel.de (M. Herde); dhuseljic@uni-kassel.de (D. Huseljic); bsick@uni-kassel.de (B. Sick); bretschneider@uni-kassel.de (U. Bretschneider); oeste-reiss@uni-kassel.de (S. Oeste-Reiß)
0000-0003-4908-122X (M. Herde); 0000-0001-6207-1494 (D. Huseljic); 0000-0001-9467-656X (B. Sick); 0000-0002-2494-0457 (U. Bretschneider); 0000-0002-6576-8841 (S. Oeste-Reiß)

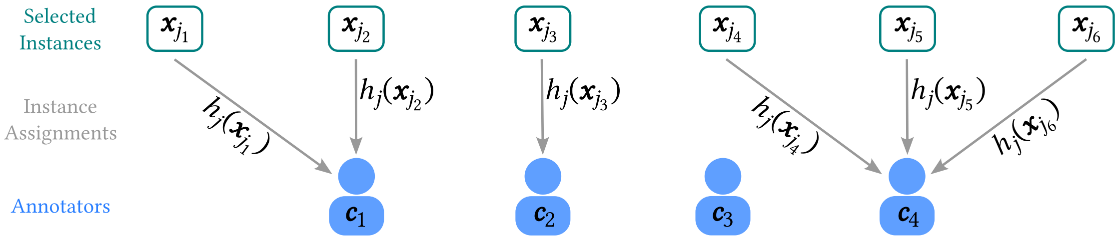

Figure 1: Exemplary iteration of a crowdworking campaign with $B = 6$ instances and $M = 4$ annotators.

2. Problem Setting

Let there be $N \in \mathbb{N}_{>0}$ instances $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)^\mathrm{T} \in \mathbb{R}^{N \times D}, D \in \mathbb{N}_{>0}$ drawn independently from an unknown probability density function $\Pr(\boldsymbol{x})$. The true class labels $\boldsymbol{y} = (y_1, \dots, y_N)^\mathrm{T} \in \{1, \dots, K\}^N, K \in \mathbb{N}_{>1}$, drawn independently from an unknown categorical distribution $\Pr(y \mid \boldsymbol{x}_n)$, are unobserved due to the lack of an omniscient annotation source. Rather, there are $M \in \mathbb{N}_{>0}$ error-prone crowdworkers $\boldsymbol{C} = (\boldsymbol{c}_1, ..., \boldsymbol{c}_M)^\mathrm{T} \in \mathbb{R}^{M \times O}, O \in \mathbb{N}_{>0}$, where $\boldsymbol{c}_m$ represents crowdworker metadata [13], e.g., educational background, interests. If such data is unavailable, each crowdworker is identified via a one-hot encoded vector, i.e., $\boldsymbol{c}_m = \boldsymbol{e}_m \in \{0, 1\}^M$. We refer to the annotation of crowdworker $\boldsymbol{c}_m$ for instance $\boldsymbol{x}_n$ as $z_{nm} \in \{1, \dots, K\} \cup \{\otimes\}$, where $z_{nm} = \otimes$ indicates an unobserved annotation. Each observed annotation z_{nm} is drawn independently from an unknown categorical distribution $\Pr(z \mid \boldsymbol{x}_n, \boldsymbol{c}_m, y_n)$. We denote annotations per instance $\boldsymbol{x}_n$ as $\boldsymbol{z}_n = (z_{n1}, ..., z_{nM})^\mathrm{T}$ and annotations of all instances as the matrix $\boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_N)^\mathrm{T}$. We define a crowdworking campaign as a process with $J \in \mathbb{N}_{>0}$ iterations. Iteration $j \in \{1, \dots, J\}$ starts with $B \in \mathbb{N}_{>0}$ selected instances $\mathcal{X}_j = \{\boldsymbol{x}_{j_1}, \dots, \boldsymbol{x}_{j_B} \mid j_1, \dots, j_B \in \{1, \dots, N\}\}$. Subsequently, we select a crowdworker for each instance by specifying instance assignments $h_j : \mathcal{X}_j \to \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_M\}$. At the end of iteration j, we update the annotations $\{z_{nm} \mid \boldsymbol{x}_n \in \mathcal{X}_j, h_j(\boldsymbol{x}_n) = \boldsymbol{c}_m\}$ to obtain the matrix $\boldsymbol{Z}_j$. Figure 1 illustrates such a crowdworking iteration. Together, the J iterations result in a sequence $\boldsymbol{Z}_0, \dots, \boldsymbol{Z}_J$ with $\boldsymbol{Z}_0$ as initial and $\boldsymbol{Z}_J$ as final annotation matrix. Given these prerequisites, we investigate two objectives for optimizing the crowdworker selection.

Objective 1: The crowdworking campaign produces a final annotation matrix minimizing the number of erroneous annotations:

$$\boldsymbol{Z}_J = \arg\min\nolimits_{\boldsymbol{Z}} \left(\sum_{n=1}^{N} \sum_{m=1}^{M} \delta(z_{nm} \neq y_n) \cdot \delta(z_{nm} \neq \otimes) \right), \tag{1}$$

where $\delta : \{\text{false}, \text{true}\} \to \{0, 1\}$ is an indicator function with $\delta(\text{false}) = 0$ and $\delta(\text{true}) = 1$.
Objective 2: The crowdworking campaign produces a final annotation matrix to learn a classification function $\hat{y} : \mathbb{R}^D \to \{1, \dots, K\}$ minimizing the expected misclassification rate:

$$\boldsymbol{Z}_J = \arg\min\nolimits_{\boldsymbol{Z}} \left(E_{\boldsymbol{x}, y} [\delta(\hat{y}(\boldsymbol{x} \mid \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{Z}) \neq y)] \right). \tag{2}$$

3. Intelligent Crowdworker Selection

We aim to select crowdworkers based on their respective performances per instance. Concretely, we interpret crowdworker performance as the probability $\Pr(z_{nm} = y_n \mid \boldsymbol{x}_n, \boldsymbol{c}_m)$ of obtaining a correct annotation. This leads to the following assignments of instances to crowdworkers:

$$h_j(\boldsymbol{x}_n) = \arg\max_{\boldsymbol{c}_m} \left(\Pr(z_{nm} = y_n \mid \boldsymbol{x}_n, \boldsymbol{c}_m)\right). \tag{3}$$

The true probabilities of correct annotations are unknown in practice. Therefore, we estimate them via *multi-annotator deep learning* (MaDL) [8], which is a state-of-the-art crowd-learning technique. MaDL uses the annotated data obtained in each successive crowdworking iteration to estimate the class probabilities of each instance and a probabilistic confusion matrix for each instance-crowdworker pair. By combining both estimates, it is then possible to approximate the annotation correctness probability $\Pr(z_{nm} = y_n \mid \boldsymbol{x}_n, \boldsymbol{c}_m)$ in Eq. 3.

4. Case Study

In this case study, we investigate the potential to optimize crowdworker selection during crowdworking campaigns. Publicly available crowdworking datasets are sparsely annotated [5], so the selection of crowdworkers is highly limited. Therefore, we rely on LETTER [14] and CIFAR10 [15] as common benchmark datasets and simulate $M = 10$ error-prone crowdworkers for each. We use standard simulation methods from literature [8] and generate varying types of crowdworkers, e.g., one adversarial crowdworker, crowdworkers specialized in certain classes, and crowdworkers specialized in certain clusters of instances. The simulated crowdworking campaign is organized into $J = 25$ iterations. Initially, each crowdworker annotates 16 randomly selected instances to obtain the initial annotation matrix $\boldsymbol{Z}_0$. In each subsequent iteration, $B = 256$ randomly selected instances are assigned to the crowdworkers for annotation. After each iteration, we train a simple multi-layer perception for the LETTER dataset and a ResNet-18 [1] for the CIFAR10 dataset. We evaluate each crowdworking campaign by quantifying the rate of obtained erroneous annotations (cf. Objective 1) and the DNN's misclassification rate on a separate test set (cf. Objective 2). For evaluation, we compare the following approaches:

- Random-DL is the baseline approach. A standard DNN is trained on the annotated instances, and the selected instances are randomly assigned to the crowdworkers.
- Random-MaDL is a more advanced approach. MaDL is trained on the annotated instances, and the selected instances are randomly assigned to the crowdworkers.
- Intelligent-MaDL is the most advanced approach. MaDL is trained on the annotated instances, and the selected instances are assigned to the crowdworkers according to Eq. 3.

Our repository at https://github.com/ies-research/intelligent-crowdworker-selection provides the approaches' hyperparameters and code. A crowdworking campaign is replicated five times for each approach and dataset. Figure 2 reports the results' means and standard deviations.

For both datasets, the approach Random-DL performs worst, indicated by the highest misclassification rate of its DNN across almost all iterations. In contrast, its erroneous annotation rate is identical to Random-MaDL (the green curve hides the blue curve) because both approaches

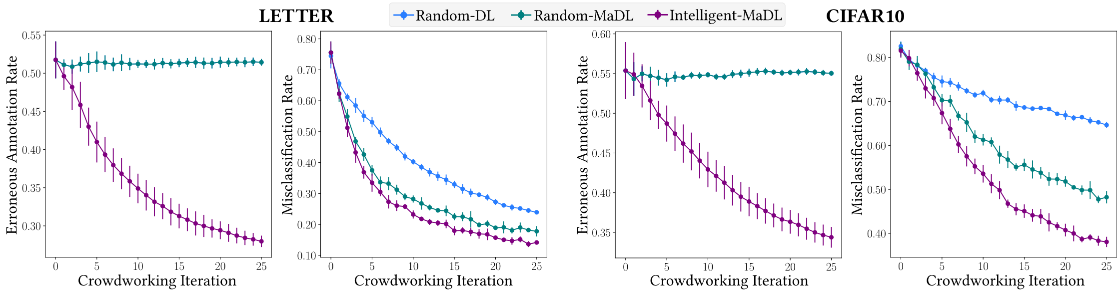

Figure 2: Results of the case study for simulated crowdworking campaigns with $J = 25$ iterations, $B = 256$ selected instances per iteration, and $M = 10$ simulated crowdworkers.

assign the instances randomly to the crowdworkers. Intelligent-MaDL consistently outperforms the other two approaches. These results confirm that MaDL improves not only the training of DNNs (lower misclassification rate of Intelligent-MaDL and Random-MaDL than Random-DL) but also the selection of crowdworkers (lowest erroneous annotation rate of Intelligent-MaDL).

5. Conclusion and Outlook

This article demonstrated the potential gains of employing a state-of-the-art crowd-learning technique during an ongoing crowdworking campaign. Our takeaways are that intelligently selecting crowdworkers reduces the number of erroneous annotations (cf. Objective 1) and improves the training of DNNs on the resulting annotated data (cf. Objective 2). Still, there are multiple future research directions to enhance the crowdworker selection further:

- Collecting metadata [13] about the crowdworkers may allow flexible and effective integration of new crowdworkers into an ongoing crowdworking campaign.
- Transferring knowledge about crowdworkers between crowdworking campaigns may improve the selection of crowdworkers for subsequent campaigns.
- Improving the uncertainty estimation [16] of crowd-learning techniques may enhance the exploration of crowdworkers' performances.
- Leveraging active learning strategies [11] to select instances intelligently may further improve the efficiency of training DNNs from crowdworking data.
- Assigning an instance to multiple crowdworkers (instead of only one crowdworker as done in Fig. 1 and Eq. 3) may better identify erroneous annotations or ambiguous instances.

For a successful deployment of intelligent crowdworker selections into actual crowdworking campaigns, we need to consider the following aspects:

- In certain settings, crowdworkers are only occasionally available, which may hinder the selection of the best crowdworker.
- Typically, the sets of instances assigned to a single crowdworker must be larger [17].
- Experiments with real-world crowdworking datasets, a larger number of annotators, and a larger number of selected instances per crowdworking iteration are necessary to validate the effectiveness of intelligent crowdworker selections.

References

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, 2016, pp. 770–778.

[2] I. Blohm, S. Zogaj, U. Bretschneider, J. M. Leimeister, How to manage crowdsourcing platforms effectively?, Calif. Manage. Rev. 60 (2018) 122–149.

[3] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, M. Allahbakhsh, Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions, ACM Comput. Surv. 51 (2018) 1–40.

[4] M. Herde, D. Huseljic, B. Sick, A. Calma, A Survey on Cost Types, Interaction Schemes, and Annotator Performance Models in Selection Algorithms for Active Learning in Classification, IEEE Access 9 (2021) 166970–166989.

[5] F. Rodrigues, F. Pereira, Deep Learning from Crowds, in: AAAI Conf. Artif. Intell., New Orleans, LA, 2018, pp. 1611–1618.

[6] H. Wei, R. Xie, L. Feng, B. Han, B. An, Deep Learning From Multiple Noisy Annotators as A Union, IEEE Trans. Neural Netw. Learn. Syst. (2022).

[7] S. Rühling Cachay, B. Boecking, A. Dubrawski, End-to-End Weak Supervision, in: Adv. Neural. Inf. Process. Syst., Virtual Conf., 2021.

[8] M. Herde, D. Huseljic, B. Sick, Multi-annotator Deep Learning: A Probabilistic Framework for Classification, arXiv:2304.02539 (2023).

[9] V. S. Sheng, J. Zhang, Machine Learning with Crowdsourcing: A Brief Summary of the Past Research and Future Directions, in: AAAI Conf. Artif. Intell, Honolulu, HI, 2019, pp. 9837–9843.

[10] B. Settles, Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[11] J. Yang, T. Drake, A. Damianou, Y. Maarek, Leveraging Crowdsourcing Data for Deep Active Learning An Application: Learning Intents in Alexa, in: World Wide Web Conf., 2018, pp. 23–32.

[12] K. Li, G. Li, Y. Wang, Y. Huang, Z. Liu, Z. Wu, CrowdRL: An End-to-End Reinforcement Learning Framework for Data Labelling, in: Int. Conf. Data Engineering, Chania, Greece, 2021, pp. 289–300.

[13] L. Zhang, R. Tanno, M. Xu, Y. Huang, K. Bronik, C. Jin, J. Jacob, Y. Zheng, L. Shao, O. Ciccarelli, et al., Learning from Multiple Annotators for Medical Image Segmentation, Pattern Recognit. (2023) 109400.

[14] P. W. Frey, D. J. Slate, Letter recognition using Holland-style adaptive classifiers, Machine Learn. 6 (1991) 161–182.

[15] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images, Master's thesis, University of Toronto, 2009.

[16] D. Huseljic, B. Sick, M. Herde, D. Kottke, Separation of Aleatoric and Epistemic Uncertainty in Deterministic Deep Neural Networks, in: Int. Conf. Pattern Recognit., Virtual Conf., 2021, pp. 9172–9179.

[17] D. E. Difallah, M. Catasta, G. Demartini, P. G. Ipeirotis, P. Cudré-Mauroux, The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk, in: World Wide Web Conf., Florence Italy, 2015, pp. 238–247.