Optimizing the Wisdom of the Crowd: Inference, Learning, and Teaching

Yao Zhou
Arizona State University
yzhou174@asu.edu

Fenglong Ma
University at Buffalo
fenglong@buffalo.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

Jingrui He
Arizona State University
jingrui.he@asu.edu

ABSTRACT

The increasing need for labeled data has brought the booming growth of crowdsourcing in a wide range of high-impact real-world applications, such as collaborative knowledge (e.g., data annotations, language translations), collective creativity (e.g., analogy mining, crowdfunding), and reverse Turing test (e.g., CAPTCHA-like systems), etc. In the context of supervised learning, crowdsourcing refers to the annotation procedure where the data items are outsourced and processed by a group of mostly unskilled online workers. Thus, the researchers or the organizations are able to collect large amount of information via the feedback of the crowd in a short time with a low cost.

Despite the wide adoption of crowdsourcing, several of its fundamental problems remain unsolved especially at the information and cognitive levels with respect to incentive design, information aggregation, and heterogeneous learning. This tutorial aims to: (1) provide a comprehensive review of recent advances in exploring the power of crowdsourcing from the perspective of optimizing the wisdom of the crowd; and (2) identify the open challenges and provide insights to the future trends in the context of human-in-the-loop learning. We believe this is an emerging and potentially high-impact topic in computational data science, which will attract both researchers and practitioners from academia and industry.

CCS CONCEPTS

• **Information systems → Crowdsourcing**;

ACM Reference Format:
Yao Zhou, Fenglong Ma, Jing Gao, and Jingrui He. 2019. Optimizing the Wisdom of the Crowd: Inference, Learning, and Teaching. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332277

1 FULL BIOGRAPHY OF THE PRESENTERS

Yao Zhou is a Ph.D. student at School of Computing, Informatics, Decision Systems Engineering, Arizona State University. He received his M.S. degree of Electrical Engineering from University of Rochester and M.S. degree of Computer Science from Oregon State University respectively. His current research interest focuses on the human-in-the-loop learning including crowdsourcing, heterogeneous learning, machine teaching, and deep learning on medical healthcare. He has published six articles in top peer-reviewed conference and journals (e.g., KDD, ICDM, SDM, IJCAI, TKDD, etc.). He has also served as a program committee members in major top conferences (e.g., ICML, NeurIPS, AAAI, IJCAI, SDM, PAKDD, etc.).

Fenglong Ma is a Ph.D. candidate of the Department of Computer Science and Engineering at University at Buffalo, The State University of New York. His research interests lie in data mining and machine learning, with an emphasis on mining health-related data. His research interests also include Crowdsourcing, Internet of Things, Social Network Mining and Security. He has published over 30 papers in top conferences and journals such as KDD, WWW, CIKM, WSDM, ICDM, IJCAI, ACL, BIBM, MobiCom and TKDE.

Dr. Jing Gao is currently an associate professor at the Department of Computer Science, University at Buffalo (UB), State University of New York. She received her Ph.D. from Department of Computer Science, University of Illinois at Urbana-Champaign in 2011. She is broadly interested in data and information analysis with a focus on truth discovery, information integration, ensemble methods, mining data streams, transfer learning and anomaly detection. She has published more than 130 papers in referred journals and conferences. She has served as program committee member of many conferences including KDD, ICDM, SDM, ECML/PKDD and CIKM.

Dr. Jingrui He is currently an associate professor at School of Computing, Informatics, Decision Systems Engineering, Arizona State University. She received her Ph.D in Machine Learning from Carnegie Mellon University in 2010. Her research focuses on heterogeneous machine learning, rare category analysis, active learning and semi-supervised learning. Dr. He is the recipient of the 2016 NSF CAREER Award, 3 times recipient of the IBM Faculty Award in 2018, 2015 and 2014 respectively, and was selected as IJCAI 2017 Early Career Spotlight. Dr. He has more than 80 publications at major conferences (e.g., IJCAI, AAAI, KDD, ICML, ICDM) and journals (e.g., TKDE, TKDD, DMKD). Her papers have been selected as Bests of the Conference by ICDM 2016, ICDM 2010, and SDM 2010.

2 TUTORIAL TOPICS

2.1 Part I: Truth Inference

In this part, we aim to answer the question: what is the ground truth label of each item that has been distributed to the crowdsourcing workers to label. Take the example of deep neural networks, which usually require million-level of labeled items (e.g., images or documents) for the model training in real-world application scenarios. The most affordable means of collecting such a large-scale data set is through crowdsourcing. Despite the wide utilization of crowdsourcing services, the frameworks of ground truth inference on these items are still not fully automatic and need human intervention to guarantee quality. To automatically estimate the trustworthy information with the data collected from online workers, truth inference

KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332277

[4, 8, 9, 11, 14, 16–19] is proposed. The challenge of truth discovery is how to estimate truths and learn accurate workers' reliability simultaneously without any supervision. Towards this goal, we will provide a comprehensive review of the existing discoveries of truth inference from the perspective of data quality in terms of label collection, worker ability modeling, item difficulty modeling, etc., and the perspective of aggregation effectiveness in terms of the optimization methods and the generative inference methods.

2.2 Part II: Learning with Crowdsourcing

Obtaining labels can be expensive and time-consuming, but unlabeled data is often abundant and easy to obtain. In terms of the labeling cost and efficiency, many learning tasks can be optimized effectively by intelligently choosing specific unlabeled instances to be labeled by an oracle, which is referred to as active learning. In practice, this setting is compromised (because this omniscient who knows ground truth doesn't exist) and the learning task could only refer to the multiple crowdsourcing workers, who have varying expertise, for label querying. By carefully chosen a set of items that are generally preferable and a group of workers to ask for queries [6, 12, 15], the items could be relabeled and have higher qualities.

Data heterogeneity, which also known as data variety, refers to the inhomogeneous properties of the data. In general, the widely accepted types of heterogeneity include task/view/oracle heterogeneity. The oracle heterogeneity is usually reflected in crowdsourcing data which collects redundant (more than one labels per item), noisy (labels could be incorrect), and possibly missing (label matrix is incomplete) labels for each item. The task heterogeneity is exhibited in the problems where multiple related predictive tasks that share commonality are jointly learned but each task has its own data set. The view heterogeneity is exhibited in the problems where the data items are characterized by different sources of features. The overall goal of heterogeneous learning is to leverage the structural information to help with model learning, e.g., aggregating the crowdsourced labels of multiple oracles [3, 18, 22], modeling the task relatedness of multiple tasks [1, 2, 22], ensure the view consistency of multiple views [13, 20] or a mixture of dual heterogeneity or triple heterogeneity.

2.3 Part III: Teaching in Crowdsourcing

In order to motivate the crowdsourcing workers to convey their knowledge more accurately, the conventional approaches are focused on designing effective incentive mechanisms with a well-designed compensation policy or a reliable privacy preserving guarantee. However, more efforts are devoted to the aspect of mechanism designing and the important fact that human beings are extremely good learning a specific concept (e.g., categorizing images, classifying text, etc) is always omitted. Besides, human beings can easily perform the concept transferring by adapting the learned concepts into new similar learning tasks. Therefore, a more effective way of utilizing crowdsourcing is by supervising the crowd to label in the form of teaching [23].

Based on the learning styles that students progress towards concept understanding, human learners can be categorized as either the sequential learners [7, 21] (who learn things in continual steps) or global learners [5, 7, 10] (who learn things in large jumps, holistically). Therefore, the state-of-the-art teaching models are usually classified into two categories based on the modeling of the learner. One category of teaching framework assumes crowdsourcing workers are global learners and their learned concepts are randomly switched in the hypotheses space. The second category of framework assumes that the workers are sequential learners who see example one at a time and make gradual progress towards becoming the experts of annotation.

ACKNOWLEDGEMENT

This work is supported by the National Science Foundation under Grant No. IIS-1552654, Grant No. IIS-1813464, IIS-1553411 and Grant No. CNS-1629888, the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-02-00, and an IBM Faculty Award. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

REFERENCES

[1] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2008. Convex multi-task feature learning. *ML* 73, 3 (2008), 243–272.

[2] Jianhui Chen, Jiayu Zhou, and Jieping Ye. 2011. Integrating low-rank and group-sparse structures for robust multi-task learning. In *KDD*. 42–50.

[3] Peng Dai, Mausam, and Daniel S. Weld. 2010. Decision-Theoretic Control of Crowd-Sourced Workflows. In *AAAI*.

[4] Pengfei Jiang, Weina Wang, Yao Zhou, Jingrui He, and Lei Ying. 2018. A Winners-Take-All Incentive Mechanism for Crowd-Powered Systems. In *NetEcon*. 3:1–3:6.

[5] Edward Johns, Oisin Mac Aodha, and Gabriel J. Brostow. 2015. Becoming the expert - interactive multi-class machine teaching. In *CVPR*. 2616–2624.

[6] Christopher H. Lin, Mausam, and Daniel S. Weld. 2016. Re-Active Learning: Active Learning with Relabeling. In *AAAI*. 1845–1852.

[7] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B. Smith, James M. Rehg, and Le Song. 2017. Iterative Machine Teaching. In *ICML*. 2149–2158.

[8] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation. In *KDD*. 745–754.

[9] Fenglong Ma, Chuishi Meng, Houping Xiao, Qi Li, Jing Gao, Lu Su, and Aidong Zhang. 2017. Unsupervised discovery of drug side-effects from heterogeneous data sources. In *KDD*. ACM, 967–976.

[10] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. 2018. Teaching Categories to Human Learners With Visual Explanations. In *CVPR*. 3820–3828.

[11] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning From Crowds. *JMLR* 11 (2010), 1297–1322.

[12] Victor S. Sheng, Foster J. Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD*. 614–622.

[13] Hua Wang, Feiping Nie, and Heng Huang. 2013. Multi-View Clustering and Feature Learning via Structured Sparsity. In *ICML*. 352–360.

[14] Houping Xiao, Jing Gao, Qi Li, Fenglong Ma, Lu Su, Yunlong Feng, and Aidong Zhang. 2016. Towards Confidence in the Truth: A Bootstrapping Based Truth Discovery Approach. In *KDD*.

[15] Yan Yan, Rómer Rosales, Glenn Fung, and Jennifer G. Dy. 2011. Active Learning from Crowds. In *ICML*. 1161–1168.

[16] Hengtong Zhang, Yaliang Li, Fenglong Ma, Jing Gao, and Lu Su. 2018. TextTruth: An Unsupervised Approach to Discover Trustworthy Information from Multi-Sourced Text Data. In *KDD*. 2729–2737.

[17] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth Inference in Crowdsourcing: Is the Problem Solved? *PVLDB* 10, 5 (2017), 541–552.

[18] Dengyong Zhou, John C. Platt, Sumit Basu, and Yi Mao. 2012. Learning from the Wisdom of Crowds by Minimax Entropy. In *NeurIPS*. 2204–2212.

[19] Yao Zhou and Jingrui He. 2016. Crowdsourcing via Tensor Augmentation and Completion. In *IJCAI*. 2435–2441.

[20] Yao Zhou and Jingrui He. 2017. A Randomized Approach for Crowdsourcing in the Presence of Multiple Views. In *ICDM*. 685–694.

[21] Yao Zhou, Arun Reddy Nelakurthi, and Jingrui He. 2018. Unlearn What You Have Learned: Adaptive Crowd Teaching with Exponentially Decayed Memory Learners. In *KDD*. 2817–2826.

[22] Yao Zhou, Lei Ying, and Jingrui He. 2017. MultiC^2: an Optimization Framework for Learning from Task and Worker Dual Heterogeneity. In *SDM*. 579–587.

[23] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. 2018. An Overview of Machine Teaching. *CoRR* abs/1801.05927 (2018).