TEXAS TECH UNIVERSITY

Edward E. Whitacre Jr. College
of Engineering™
Computer Science

"Data Quality and Learning with Crowdsourcing"

Shengli Sheng

Prospective Faculty Candidate

University of Central Arkansas

Thursday, February 28, 2019
2:30 p.m.
ECE, Room 217

Abstract

Crowdsourcing systems provide convenient platforms to collect human intelligence for a variety of tasks (e.g., labeling objects) from a vast pool of independent workers (a crowd). Compared with traditional expert labeling methods, crowdsourcing is obviously more efficient and cost-effective, but the quality of a single labeler cannot be guaranteed. In taking advantage of the low cost of crowdsourcing, it is common to obtain multiple labels per object (i.e., repeated labeling) from the crowd. In this talk, we outline our research on crowdsourcing from three aspects: (1) crowdsourcing mechanisms, specifically on repeated labeling strategies; (2) ground truth inference, specifically on noise correction after inference and biased wisdom of the crowd; and (3) learning from crowdsourced data. We first present repeated-labeling strategies of increasing complexity to obtain multiple labels. Repeatedly labeling a carefully chosen set of points is generally preferable. A robust technique that combines different notions of uncertainty to select data points for more labels is recommended. Recent research on crowdsourcing focuses on deriving an integrated label from multiple noisy labels via expectation-maximization based (EM-based) ground truth inference. We present a novel framework that introduces noise correction techniques to further improve the label quality of the integrated labels obtained after ground truth inference. We further show that biased labeling is a systematic tendency. State-of-the-art ground truth inference algorithms cannot handle the biased labeling issue very well. Our simple consensus algorithm performs much better. Finally, we present pairwise solutions for maximizing the utility of multiple noisy labels for learning. Pairwise solutions can completely avoid the potential bias introduced in ground truth inference. They have both sides (potential correct and incorrect/noisy information) considered, so that they have very good performance whenever there are a few or many labels available.

Bio

Victor S. Sheng received the M.Sc. degree from the University of New Brunswick, Fredericton, NB, Canada, and the Ph.D. degree from the University of Western Ontario, London, ON, Canada, both in computer science, in 2003 and 2007, respectively. He is an Associate Professor of computer science and the Founding Director of Data Analytics Laboratory at University of Central Arkansas. After receiving the Ph.D. degree, he was an Associate Research Scientist and NSERC Postdoctoral Fellow in information systems with the Stern Business School at New York University. His research interests include data mining, machine learning, crowdsourcing, and related applications in business, industry, medical informatics, and software engineering. He has published more than 140 research papers in conferences and journals of machine learning and data mining. Most papers are published in top journals and conferences in data science, such as PAMI, TNNLS, TKDE, JMLR, AAAI, KDD, IJCAI, and ACMMM. Prof. Sheng is a senior member of IEEE. He is a conference organizer for several conferences, and an editorial board member for several journals. He also is a SPC and PC member for many international conferences (such as IJCAI, AAAI, and KDD) and a reviewer of more than twenty international journals (such as PAMI, TNNLS, TKDE, and JMLR). He was the recipient of the Best Paper Award Runner Up from KDD'08, the Best Paper Award from ICDM'11, the Best Student Paper Award Finalist from WISE'15, and the Best Paper Award from ICCCS'18.