Similarity Learning
in the Era of Big Data

Shiyu Chang

Shiyu Chang is a research staff member at the IBM Thomas J. Watson Research Center. He recently obtained his Ph.D. from the University of Illinois at Urbana-Champaign (UIUC) under the supervision of Prof. Thomas S. Huang. Chang has a wide range of research interests in data explorations and analytics at large-scale. Specifically, his current research directions lie on developing novel machine learning algorithms to solve complex computational tasks in the real-world. Chang received his B.S. degree at UIUC in 2011 with the highest university honor (Bronze Tablet Award). He graduated from the Department of Electrical and Computer Engineering at UIUC and obtained his M.S. degree in 2014. He is a recipient of the Thomas and Margaret Huang Award in 2016 and the Kodak Fellowship Award in 2014. Most of Chang’s research has been published in top data mining, computer vision, and artificial intelligent venues including SIGKDD, WWW, CVPR, WSDM, ICDM, SDM, IJCAI etc. The paper “Factorized Similarity Learning in Networks” has been selected as the best student paper in ICDM 2014.

Abstract:

The notion of machines that can learn has caught imaginations since the days of the early computer. In recent years, as we face burgeoning amounts of data around us that no human mind can process, machines that can learn to automatically find insights from such vast amounts of data have become a growing necessity. The field of machine learning is a modern marriage between computer science and statistics driven by tremendous industrial demands. The soul behind many applications is based on the so-called "similarity learning”. Learning similarities is often used as a subroutine in important data mining and machine learning tasks. For example, recommender systems utilize the learned metric to measure the relevance of the candidate items to target users. Applications of this approach also exist in the context of clinical decision support, search, and retrieval settings. However, the three-V (volume, variety, and velocity) natures of big data make learning similarity for pattern discovery and data analysis create new challenges: How to reveal the truth from massive unlabeled data? How to handle data with multimodality? What if the data consist network structures? Does temporal dynamic effect the process of decision-making? For example, in clinical decision making, doctors retrieve the most similar clinical pathway for auxiliary diagnosis. However, the sheer volume and complexity of the data present major barriers toward their translation into effective clinical actions. In this talk, Chang will illustrate some of these challenges with examples from his works on foundations of similarity learning. He will show that with judicious design, together with rigorous mathematics for learning similarities, we are able to make various kinds of impact on society and uncover surprising natural and social phenomena.