UNIVERSITY OF CALIFORNIA MERCED | SCHOOL OF ENGINEERING

Electrical Engineering and Computer Science Seminar

# Towards automated data science: automating feature extraction and feature selection

**Dr. Yanjie Fu**

**Friday, February 21, 2020**

**12:00 PM – 1:20 PM in COB 114**

Faculty Host: Prof. Wan Du


**Abstract**

The end to end automation of deep learning has been successful in extremely data-intensive computing, such as images, videos, audios. However, are we really making much progress? The best paper of ACM RecSys Conference 2019 pointed out that: despite amazing accuracy, deep learning suffers from low reproducibility and high effort of model tuning. On the contrary, classic data mining pipeline (data preprocessing, feature extraction, feature selection, predictive modeling) has been proved to be effective, step by step testable, and of low tuning effort. Can we integrate the automation with data mining pipelines to achieve a balance between automation and step by step testable low-effort reproducibility? In this talk, I will introduce the motivation and significance of our overall goal to automate a classical data mining pipeline. Then, I will discuss two preliminary studies: (i) automated characterization (AKA feature extraction) of spatial-temporal graphs, (2) automated feature selection via multi-agent reinforcement learning. Finally, I will conclude my talk and present our future work.



soegrads@ucmerced.edu



**Dr. Yanjie Fu**
**University of Central Florida**


**Biography**

Dr. Yanjie Fu is an assistant professor in the Department of Computer Science at the University of Central Florida. He received his Ph.D. degree from Rutgers, the State University of New Jersey in 2016, the B.E. degree from University of Science and Technology of China in 2008, and the M.E. degree from Chinese Academy of Sciences in 2011. His research interests include data mining and big data analytics. He has research experience in industry research labs, such as Microsoft Research Asia and IBM Thomas J. Watson Research Center. He has published prolifically in refereed journals and conference proceedings, such as IEEE TKDE, ACM TKDD, IEEE TMC, ACM TIST, ACM SIGKDD, AAAI, and IJCAI. His general research goal is to develop algorithms and tools to answer: (1) how data mining approaches alleviate information heterogeneity, dynamics, and unstructureness; and (2) what role modeling structures play in exploring the correlations among structure-buried data. He aims to strategically reformulate the problem of modeling structure-buried data into a new machine learning task, and develops robust algorithms for the task. He also investigates how the developed algorithms can be applied for real world problems, including mobile, transportation, power, IoT, and education analytics. His recent research focuses on mining spatial-temporal graph-structured behavioral data, and automated data science.