Pattern Discovery and Large-Scale Data mining on cosmological datasets

Doris Jung Lin Lee (UC Berkeley), Robert J. Brunner (UIUC)

With next-generation telescopes capturing tens of TB/night of observational data, the role of scalable and efficient data analysis methods have become central to the knowledge discovery process in Astronomy. The diversity and scale of astronomical datasets also presents challenging research problems to the data mining and machine learning community. This poster describes three projects highlighting our recent work in these areas: 1) Current state-of-the-art ML algorithms (SVM, LDA, DNNs) are capable of classifying galaxy morphology at above 90 percent accuracy, but their results reflect the inherent errors due to human classifiers. We propose a scalable, hybrid technique that integrates active learning in crowdsourcing citizen science platforms for improving the data quality of the training labels. 2) We developed a recursive, source-finding algorithm that automatically corrects for positional inaccuracies in outdated astronomical catalogs. By applying this technique to imaging data from two different sky survey, we recovered all 23,011 sources in a widely used astronomical catalog. 3) Traditional friends-of-friends algorithms and density-estimation methods designed for halo-finding are not only computationally intensive, but especially problematic for detecting substructures within haloes. We explore non-parametric, unsupervised methods for finding haloes in the Dark Sky Simulation, a 34TB N-body simulation containing trillions of particles.