2015 NSF IIS-GEO Workshop Statement

Yan Liu (University of Southern California)

Recent development in machine learning In the past few decades, we have witnessed major breakthrough and significant progress in machine learning and data mining research. Many novel algorithms have been developed and proven to be effective in challenging data analysis and prediction tasks, such as image analysis, speech recognition, question answering, machine translation and so on. Here I will give three examples of active research directions in machine learning that are closely related to geoscience:
(1) *Large-scale time series and spatial-temporal data analysis*: Before the era of big data, time series and spatial time series analysis were mostly pursued by researchers in statistics, economics, and finance, focusing on forecasting (i.e., prediction) tasks. With the development of sensor technologies, more and more large-scale time series and spatial-temporal data are available. As a result, many learning, mining and modeling problems natural emerge in scientific domains (including geoscience), such as frequent temporal pattern mining, temporal dependence discovery, trajectory mining, change point detection (or anomaly detection), and variations of state-space models. In recent years, active research efforts have been devoted to scaling up machine learning and data mining models, developing online update algorithms and seeking interpretations of the results.
(2) *Causal inference from big data*: causal inference, i.e., identifying the causes of an event or phenomena from observational data, has always been sought after by researchers in many disciplines, ranging from SEM framework by Pearl, Rubin causal models by Rubin, to PC algorithms by Spirtes. Recent research activities in causal inference have led to interesting breakthrough on the practical side of causal inference, such as nonlinearity, additive noise models, transportability, causal priors (such as temporal information), and so on. There is no doubt that these developments can help solve interesting scientific problems in geoscience.
3) *Probabilistic programming*: In the process of scientific discovery, scientists have accumulated a large amount of knowledge around the earth. Even though large-scale data have been collected, leveraging domain knowledge will usually help improve the effectiveness of inference and learning, leading to more robust and interpretable models. Recent successes in probabilistic programming systems or probabilistic logic systems provide easy platforms for us to incorporate domain knowledge into statistical models. The major challenges are how to effectively represent domain knowledge and how to efficiently make inferences in these systems.

Important research trends in IIS for Geoscience The intersection between machine learning, data mining and geoscience is an active research field. Many communities, workshops, and special tracks in major conferences have been developed in the past few years. For example, the climate informatics community (`http://www.climateinformatics.org`) draws researchers in machine learning, data mining, and statistics to work with climate scientists on important climate-related problems. Its associated annual workshop, i.e., International Climate Informatics Workshop, is held in NCAR, Boulder, CO and reaches its fifth anniversary in 2015 (also supported by NSF). The workshop has grown significantly over the years in terms of the number of attendee and the number of abstract submissions (The statistics from last year is over 70 attendee, over 50 submission and 40 accepted abstracts). In addition, AAAI and IJCAI both have hosted special track in sustainability for researchers to publish novel work on machine learning, artificial intelligence and data mining around earth science.