NSF Discovery Informatics Workshop (DIW) Statement

Yan Liu (University of Southern California)

January 25, 2012

The rapidly grown data in scientific field have fundamentally reshaped the ways we learn about the world and conduct scientific research. Many challenges must be addressed in the areas of information sciences, intelligent systems, and human-computer interaction to analyze the data and facilitate scientific discovery. As a researcher in machine learning and data mining, I would like to share my personal views on two key machine learning challenges in discovery informatics and discuss potential solutions, including *temporal dependency analysis* and *learning in big data*.

Temporal Dependency Analysis for Scientific Discovery With the technology breakthrough in many scientific domains, we are able to collect a large number of valuable time series observations. For example, in the field of biology, time series microarray data recording the gene expression levels under different treatments over time are available to study their regulatory functions and cell-cycles; in climate science, climate and meteorological data, such as temperature, solar radiation, and carbon-dioxide concentration, are collected over the years to help understand the climate system and attribution factors of global warming; in social science, social media contents can be collected over time to examine information diffusion and social influence.

To scientific researchers, learning temporal dependencies between time series variables is extremely valuable for providing insights into the scientific questions that we are interested in answering, e.g., which set of genes down-regulated CDK2 kinases, and thus promotes both cell cycle G1/S and G2/M transitions? What climate forcing agents causes the frequent extreme heat or draught in climate system? Which Facebook or Twitter accounts are most influential to a target user for marketing? As we can see, the answers to these questions directly generate scientific hypotheses, which can significantly reduce the searching space for scientific discovery.

Up to now, a number of approaches have been investigated to define correlations or causalities for time series data, such as autocorrelation, cross-correlations, randomization test, phase slope index, transfer entropy etc. Most approaches are based on statistical significance tests, which are time-consuming: the complexity grows exponentially with the number of features involved. Our recent work investigates a novel model, namely *Granger graphical models* [1], as an effective, robust, yet scalable solution to learning temporal dependencies in time series data. The model automatically discovers temporal dependence, in the semantics of *Granger causality*[2], from high-dimensional time series data via penalized regression. We have successfully applied our model to applications in system biology (e.g., gene-regulatory network discovery from time series microarray data) [5, 3] and climate data analysis (e.g. climate change attribution) [6, 4]. We have also identified several other scientific applications, such as astronomy, marine biology, health care, power grid, business analytics and so on, and successfully set up collaborations with experts in each domain. The challenges in analyzing real application data also push the frontiers of machine learning research. Some promising directions include nonlinear time series analysis, causality analysis with hidden variables, irregular time series analysis, relational time series analysis, and extreme-value time series analysis.

Mining and Learning in Big Data One of the biggest challenges for many *advanced* machine learning models to be applicable in real applications is that they are not scalable to large-scale data. Recently, there have been many efforts to address this issue. These efforts can be loosely categorized into three communities:

One is the machine learning community, who develop general-purpose parallel machine learning packages with broad users and applications in mind. The potential disadvantage is that usually users have to acquire significant amount of machine learning knowledge as well as coding skill to successfully utilize these packages in their real applications. Furthermore, the general-purpose packages are usually not optimized for specific problems. Therefore to make these packages successful, machine learning experts may need to work closely with domain experts in at least one application and take the time to immerse with the scientific community by providing tutorials and demos;

The other community is researchers in interdisciplinary areas, who utilize machine learning and data mining techniques to solve specific problems of interest (e.g. computational biology). The dilemma that researchers in this field are confronted with is that they are very interested in applying advanced machine learning techniques to solve the problems while most often than not, simple algorithms perform quite well and are more robust. For some tasks that cannot be addressed by standard algorithms, they have to develop their own models with unique tailoring which general-purpose packages cannot achieve (furthermore, the amount of time and effort on getting familiar with the packages in order to make changes on the code is so much that it could be the case that researchers would prefer to develop their own package).

The third community is the scientific computing community. It is a very mature field concerned with constructing mathematical models and quantitative analysis techniques to analyze and solve scientific problems. There already exist very good infrastructures to support large-scale data (e.g. supercomputer centers) and researchers in this field have rich experience in handling big data. However, most efforts are centered around the applications of computer simulation or other forms of computation in applying known techniques.

There are many other excellent efforts on big data learning in addition to the three we discussed above. Nevertheless, the main argument we aim to make is as follows: researchers in different communities have independently made significant efforts on analyzing large-scale scientific data in the past; cross-community activities have already happen, but the level of integration and collaboration is far behind what is needed for truly facilitating scientific discovery. The solution to this problem is definitely not an easy path since each community has their own problem of interest. Education and interdisciplinary programs could help significantly.

References

[1] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical Granger methods. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-07)*, 2007.

[2] C. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[3] Y. Liu, A. Niculescu-Mizil, A. Lozano, and Y. Lu. Temporal graphical models for cross-species gene regulatory network discovery. In *International conference on Computational Systems Bioinformatics (CSB 2010)*, 2010.

[4] Y. Liu, A. Niculescu-Mizil, A. C. Lozano, and Y. Lu. Learning temporal causal graphs for relational time-series analysis. In *Proc. of ICML'10*, pages 687–694, 2010.

[5] A. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical Granger modeling for gene expression regulatory networks discovery. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB-09)*, 2009.

[6] A. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-09)*, 2009.