Exascale Data Mining for Visual Data Analytics

Dr. Wei Xu, Dr. Dantong Yu, and Dr. Shinjae Yoo and, CSC BNL {xuw, dtyu, sjyoo}@bnl.gov

"Big data" from experiments, simulations, the literature, and social networks offer significant challenges of Variety, Volume, and Velocity (3-Vs) in data management and sharing, data mining, and knowledge discovery, and current computing paradigms. To resolve these programs, a holistic approach must be co-designed from the low-level hardware accelerated computing, to the top-level application-driven knowledge discovery and human-computer interaction.

Large experiments facilities, for example, NSLS-II, LCLS,, along with new sequencing technology innovations system biology, healthcare generate petabytes/exabytes of image data, simulation output, and text data. Heterogeneous data must be reconstructed, analyzed, mined, integrated, to enable large scale knowledge discovery. Status-quo data analysis only considers one technique with one parameter setting, applied in a single step. Scientific research is overwhelmed by the sheer volume of data and their complexity, as measured in three aspects (3-Vs): Variety, volume and velocity. Visual analytics offers a convenient way to involve the user in problem-solving through visualization. For a large ecosystem body of data, this approach would be extraordinarily helpful in extending human experience/intelligence, enabling cognitive operations, enhancing the hypothesis generation, and testing them. Using data mining techniques would facilitate the detection of anomalies, causality or correlation in data, and eventually decision making via a human factor-aware visual analytic user interface. Adopting the four areas described as follows will greatly improve the analysis and understanding of data ecosystems:

Area 1: Interactive efficient parameter tuning assisted by visual data analytics

Nearly all the data mining and machine learning methods have parameters that must be tuned precisely; inappropriately setting the parameters will lead to misinterpreting data. With the 3-V challenges, it is very difficult to find an optimal parameter setting to fit more than one dataset. Therefore, we need 1) a learning model that simultaneously will identify multiple parameter settings , and 2) a system wherein if a data analyst selects and tunes the specific parameter settings, then given the current models, the input from domain experts, in terms of parameter tuning and setting can be efficiently and progressively incorporated into new models.

Area 2: Approximated online adaptive learning

It is highly desirable to obtain the approximate outcomes rapidly for processing the data-driven learning models. Obtaining the approximate results quickly then can subsequently steer the user to more expensive, complex, yet comprehensive analysis. Especially, online streaming learning methods are among those of the greatest interest due to their one-pass learning capabilities. On top of learning from online streaming data, the analyst can provide feedbacks based on the intermediate results to the learning model which subsequently can adapt to these feedbacks (in other words, adaptive learning). To cope with the 3V challenges, especially high velocity, visual analytics can significantly enhance approximated online adaptive learning by incorporating domain experts into the learning loop.

Area 3: Robust un-/semi-supervised learning

It is difficult to annotate a large volume of data during supervised learning due to the excessive cost of human involvement. Unsupervised or semi-supervised learning becomes critical to visual data analytics. However, compared to supervised learning, these two models tend to be highly sensitive to tune parameter when there is a lack of appropriate annotations and users' feedback. Even with incomplete data analysis that only have partial results available, semi-supervised learning models can effectively can help analysts in finding targets progressively and quickly. We recommend robust unsupervised and supervised learning techniques to capture recurrent data

events and unseen patterns easily and effectively that might be candidates for annotation and parameter tuning. Visual data analysis is the bridge to integrate these two types of learning.

Area 4: Algorithm Acceleration with state-of-the-art hardware

In response to the anticipated high volume of data generated by advanced observational platforms and fine-grain model simulations, the models and methodology are often data-/computing intensive. It is highly desirable to extensively use state-of-the-art computational technologies, such as high-performance parallel processing, GPGPU and the Intel Phi coprocessor to accelerate algorithm modules, building blocking, dependency libraries and also to cope with the challenges of data volume and velocity.

In pursuit of the concepts discussed above, we offer a few suggestions for consideration:

- There needs to be continued- and increased- investment in the research on and development of basic algorithm research in data mining and knowledge discovery.
- Data mining and visualization needs to be integrated to cross-validate the results of data mining by human involvement. However, directly using existing visual analytic techniques can hardly satisfy the demands for analyses. Customization is needed when developing visual analytics tool for specific applications. The development usually includes the collaboration of experts from related multiple disciplines. There are many cutting-edge publications as showcases of the customization process with current visual analytics techniques. Undoubtedly, more investment is needed to transfer these research outcomes into practical system and foster new analysis algorithm to be developed.
- On-line data sampling and streaming analysis will be critical as data must be analyzed before most of it is archived and analysis of it becomes difficult.
- Data integration and the analysis of fused data must be supported to reveal a comprehensive picture of science phenomena as the same data samples are measured, and evaluated under various circumstances.

Wei Xu received her Ph.D. degree in Computer Science from Stony Brook University, USA in 2012. She joined Brookhaven National Lab (BNL) in 2013. Her research interests include medical imaging, tomography, visualization, visual analytics, GPGPUs, machine learning and workflow systems. She has published papers in leading technical journals and conferences and served as committee member and reviewers for top medical imaging and visualization journals and conferences.

Dantong Yu is a research scientist and group leader at the computational science center. His research interests include HPC, data mining and science-driven knowledge discovery. He has published more than 50 papers in leading CS journal and conference proceedings.

Shinjae Yoo is a research scientist and group leader at the computational science center. His research interests include bioinformatics, data mining and machine learning. He has published work on data mining and machine learning.

References

- Daniel A. Keim, Jörn Kohlhammer, Geoffrey Ellis, Florian Mansmann, “Mastering The Information Age – Solving Problems with Visual Analytics”, Florian Mansmann.
- H. Huang, S. Yoo, H. Qin, and D. Yu, Physics-based Anomaly Detection Defined on Manifold Space, Accepted to ACM Transactions on Knowledge Discovery from Data, 2014 (TKDD).
- Huang, H., Yoo, S., Yu, D., and Qin, H., “Noise-Resistant Unsupervised Feature Selection via Multi- Perspective Correlations”, regular paper, IEEE International Conference on Data Mining (ICDM), December 2014.
- Huang, H., Yoo, S., Yu, D., and Qin, H., “Diverse Power Iteration Embeddings and Its Applications”, regular paper, IEEE International Conference on Data Mining (ICDM),