Mining science data

Chandrika Kamath
Lawrence Livermore National Laboratory, Livermore, CA 94551
E-mail: kamath2@llnl.gov

Abstract.
The data from scientific simulations, observations, and experiments are now being measured in terabytes and will soon reach the petabyte regime. The size of the data, as well as its complexity, make it difficult to find useful information in the data. This is of course disconcerting to scientists who wonder about the science still undiscovered in the data. The Sapphire scientific data mining project is addressing this concern by applying data mining techniques to problems ranging in size from a few megabytes to a hundred terabytes in a variety of domains. In this paper, we briefly describe our work in several applications, including the identification of key features for edge harmonic oscillations in the DIII-D tokamak, classification of orbits in a Poincaré plot, and tracking of features of interest in experimental images.

1. Introduction
Advances in technology are enabling us to collect data far more rapidly than we can analyze it. This is especially true in many scientific domains, where vast amounts of data are being collected through observations, experiments, and simulations. These datasets are not only massive, being measured in terabytes and beyond, they are also very complex. As a result, data mining techniques are becoming invaluable in helping scientists find useful information in their data.

There are several aspects of science data that make it a challenge to mine such data. Science data is usually available in the raw form such as noisy images, as structured or unstructured mesh data with physical variables at each mesh point, or as the output of different sensors. As a result, a substantial amount of pre-processing is needed to bring this raw data into a form suitable for the detection of patterns or useful information. Science data can also be high-dimensional, where multiple features or characteristics are used to represent the objects or structures of interest. It often has both a spatial and a temporal component. In addition, the validation of the patterns (or information) identified by data mining is an important, though often overlooked, post-processing step in mining scientific data.

To address these issues, the Sapphire project (http://www.llnl.gov/casc/sapphire) in scientific data mining at Lawrence Livermore National Laboratory, is conducting research in robust, accurate, and scalable algorithms for all phases of the scientific data mining process, converting the research into software, and applying the software to various data analysis problems. Our research and software are motivated by the needs of our applications. We view data mining as an interactive and iterative process involving data pre-processing, search for patterns, knowledge evaluation, and possible refinement of the process based on input from domain experts or feedback from one of the steps. The pre-processing of the data can include tasks such as removing...
the noise in the data, sampling or multi-resolution analysis to reduce its size, identifying the objects in the data, extracting the features characterizing the objects, and reducing the number of features to keep only the relevant ones.

To address the diverse needs of various applications, our system architecture is composed of separate modules for the various tasks in the data mining process. For each task, such as de-noising of data or dimension reduction, we support several algorithms as the choice of the algorithm often depends on the data and the problem being analyzed. Further, the parameters for each algorithm can be tuned appropriately as required by a problem. Figure 1 describes the Sapphire system architecture, with the items in blue indicating the compute intensive parts, which are the focus of our research. The domain specific parts, highlighted in yellow, include the reading, writing and display of the input data, as well as the display of the information extracted from the data. We support various data formats including FITS, PGM, View, and netCDF, converting them to our internal format for processing. In these tasks, we make extensive use of public-domain software. Any intermediate data extracted from the raw images or meshes, is stored in the public-domain RDB software [1], which essentially acts as a simple data store.

Sapphire software has been used for the analysis of data in a variety of problems ranging from the analysis of bubbles and spikes in Rayleigh-Taylor instability to the comparison of simulations to experiments for Richtmyer-Meshkov instability, the detection of human settlements in satellite imagery, the classification of bent-double galaxies, the detection and tracking of moving objects in video, the separation of signals in climate data, and information retrieval. Next, I briefly describe three applications which have been, or are being, analyzed using Sapphire software.

2. Analysis of Poincaré plots

In collaboration with Abraham Bagherjeiran, Erick Cantú-Paz, Siddharth Manay, Dale Slone, Neil Pumphrey, Don Monticello, and Scott Klasky.

An important step in the quest for low-cost fusion power is the ability to perform and analyze experiments in prototype fusion reactors such as the National Compact Stellarator Experiment at the Princeton Plasma Physics Laboratory (PPPL). These devices allow physicists to perform magnetic confinement experiments which determine the best shape for the hot
reacting plasma and the magnetic fields necessary to hold it in place. In addition, advances in computational resources have made possible the computational simulation of these experiments in three dimensions over time. This allows the physicists to design new reactors and select the parameters to be used in experiments. The experimental results are, in turn, used to validate the simulations. Thus, the analysis of data from both simulations and experiments is a key step in the understanding and development of fusion reactors. One of the tasks in the analysis is the classification and characterization of orbits in Poincaré plots generated by the particles in a fusion reactor as they move within the toroidal device.

As a particle moves around the torus, it will trace out a three-dimensional trajectory over time. Consider a plane intersecting the torus perpendicular to the magnetic axis, that is, a vertical slice through the torus. Let a point in this plane be the intersection of the trajectory of the particle with the plane as it starts to move through the torus. After it completes one round through the torus, it will likely intersect the plane at a different point. The intersections of this trajectory with the plane form an orbit. Depending on the shape of the orbit, it can be assigned a class label. Figure 2 depicts three different orbits: a quasi-periodic orbit, an island chain, and a separatrix.

Our goal in this problem is to classify an orbit into one of the three categories using the data available from simulations, where each orbit is composed of a series of points. This can be difficult when an orbit is described by few points, or an orbit of one type visually appears as another due to the location and number of points. We are currently investigating two approaches to address this problem. The first is based on converting the data into polar coordinates, extracting features based on the distribution of the points, and then using simple rules on these features to identify the class of the orbit. The second approach is graph-based and considers the graph corresponding to the minimal spanning tree of the points, extracts features from the graph, and uses the features in simple rules for classification. Our early work [2] showed that by using machine learning techniques on the graph-based features, we could substantially improve the accuracy of the approach.

We are currently testing and enhancing our software for the classification of the Poincaré plots. Future work in this area will include the extraction of characteristics such as the number of islands in an island chain, the width of the islands, and the width of the lobes of the separatrix.

3. Identification of features for edge-harmonic oscillations

In collaboration with Erick Cantú-Paz, Keith Burrell, and Mike Walker.

In some problems, instead of predicting the class of a object such as an orbit, we are interested in finding out which features are important to the phenomena being studied. For example, in a tokamak, the preferred mode of operation is the high-confinement mode (H-mode), which comes at a significant cost due to effects of edge localized modes (ELMs). ELMs cause rapid erosion of

![Figure 2. Sample orbits. (a) Quasi-periodic, (b) Island Chain, and (c) Separatrix.](image-url)
some components in tokamaks and giant ELMs can destroy other critical components. Recently, a “quiescent H-mode” of operation has been observed in the DIII-D tokamak operated by General Atomics. This quiescent operation is important because it has no ELMs. Further, physicists have found that a phenomena known as the edge harmonic oscillation (EHO) is associated with the quiescent H-mode. They are able to identify these EHOs both visually and using simple rules derived from the visual analysis, but they have been unable to explain when or why EHOs appear.

Our goal in this analysis was to identify which of the variables being measured by different sensors were relevant to the presence of EHOs. First, with input from the physicists, we extracted the values of 37 candidate variables that describe approximately 700 experiments, each lasting about 6 seconds. Each 50 ms time window of each experiment received a binary label (high/low EHO-ness) using the program that detects EHOs. Next, we preprocessed the data to discard the time windows that contain at least one missing value. This could be due to a sensor being inactive, or data from all sensors not being sampled at the same rate. In addition, a visual examination of box-plots and histograms revealed that the data contained many outliers. Using the median value of each variable in each time window eliminated some outliers, but since many still remained, we decided to eliminate the time windows that contained at least one variable in the top or bottom percentile of its range. After the preprocessing, our training set consisted of 41818 instances.

Figure 3 presents the error rates of a naive Bayes classifier trained on increasingly large feature subsets [3]. We found that the Principal Component Analysis (PCA) filter produced a compact feature subset that results in the lowest classification error of 17.3%. Although this error is not notably smaller than the error obtained with all the features (20.9%), the fact that very few features are necessary to explain the presence of EHOs is interesting. We also found that there is significant overlap between the top ten features ranked by the different methods, except for the PCA filter which selects features that the other methods rank lower. Six features were ranked in the top ten by four filters, and an additional two were ranked in the top ten by

![Figure 3. Error rates varying with the number of features using the fusion data. The large dots represent the rankings of the random noise feature.](image-url)
Figure 4. (a) Original experimental image from NSTX and (b) after denoising. It is unclear if the bright line to the left of the bright blob is noise or signal.

Figure 5. The large change in intensity at the center between two consecutive images from NSTX indicates that optical flow techniques are inappropriate for tracking.

three filters. This reduced subset of features was given to the physicists for further investigation.

4. Feature tracking algorithms
In collaboration with Cyrus Harrison, Nicole Love, and Stewart Zweben

In many datasets, both from computer simulations and experiments, there is an interest in identifying and tracking structures of interest. These structures could be blobs in plasma, bubbles and spikes in fluids, or particles in a simulation. As the problem of tracking occurs in many domains, we are developing a suite of algorithms which can be applied as appropriate to a given dataset. There is clearly a need for a variety of algorithms as a single algorithm is unlikely to work for all problems for all cases. For example, consider the problem of tracking coherent structures in plasma, such as the ones seen in images taken using the gas-puff imaging system at the National Spherical Torus experiment (NSTX) (http://www.pppl.gov/~szweben/NSTX04/NSTX_04.html) at PPPL. These experimental images are noisy (Figure 4a), and after denoising (Figure 4b), it is unclear if the vertical streak to the left of the bright blob is part of the signal or the noise. Further, two consecutive frames (Figure 5) from a sequence show that the intensity at a point could change rapidly, making techniques such as optical flow, inapplicable for tracking the blobs. However, if the intensity change is gradual over time, optical flow techniques can be used to obtain an estimate of the motion of the plasma.

Acknowledgments
The Sapphire team gratefully acknowledges the domain scientists for sharing their data and expertise with us. The work on the three applications presented in this paper was partially funded by the DOE Office of Science through the SciDAC-1 program. In addition to SciDAC, the Sapphire project is funded by the DOE NNSA ASC program and the LDRD program at Lawrence Livermore National Laboratory.

UCRL-CONF-222246. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

References
[1] Hobbs, W. A relational database management system, 1993. Available at ftp://ftp.rand.org/pub/RDB-hobbs/.
[2] Bagherjeiran, A. and Kamath, C., Graph-based Methods for Orbit Classification, In Proceedings of the Sixth SIAM International Conference on Data Mining, pp 574-578, Bethesda, MD, 2006.
[3] Cantó-Paz, E., Newsam, S., and Kamath, C., Feature Selection in Scientific Applications, In Proceedings, ACM International Conference on Knowledge Discovery and Data Mining, pp 788-793, Seattle, WA, 2004.