UCRL-ID-130245

Scalable Pattern Recognition for Large-Scale Scientific Data Mining

C. Kamath
R. Musick

March 23, 1998

Lawrence Livermore National Laboratory

This is an informal report intended primarily for internal or limited external distribution. The opinions and conclusions stated are those of the author and may or may not be those of the Laboratory.

Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract W-7405-ENG-48.

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes

This report has been reproduced
directly from the best available copy.

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information
P.O Box 62, Oak Ridge, TN 37831
Prices available from (423) 576-8401

Available to the public from the
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd.,
Springfield, VA 22161

A White Paper Proposing a Project

in

Scalable Pattern Recognition for Large-Scale Scientific Data Mining

Chandrika Kamath
Ron Musick
(email: kamath2, musick2@llnl.gov)

March 23, 1998 at 12:45

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

1 Goals

We are proposing a research effort within CASC in the area of data-mining, in particular, pattern recognition for large, multi-dimensional, scientific data sets. These techniques are extremely useful in improving the way scientists interact with data to construct and validate computational models of physical events. This project would:

- Enhance the ability of pattern recognition algorithms to accurately and efficiently model complex phenomena by integrating ideas from scientific computing (SVD and wavelets) and soft computing (neural nets and decision trees).
- Scale pattern recognition techniques to large scientific data sets by applying techniques from high performance computing, such as efficient algorithms and parallel processing.
- Support the use of interactive exploration of large data sets by employing techniques such as dimension reduction, random sampling, and multi-resolution data.
- Enable the user to control the tradeoff between computational effort, and the accuracy of the models derived from the pattern recognition process. The application and extension of techniques from stochastic modeling and machine learning will provide a mathematical basis for this work.

2 Motivation

Our ability to generate data far outstrips our ability to explore and understand it. The true value of this data lies not in its final size or complexity, but rather in our ability to exploit the data to achieve scientific goals.

The data generated by programs such as ASCI have such a large scale that it is impractical to manually analyze, explore, and understand it. As a result, useful information is overlooked, and the potential benefits of increased computational and data gathering capabilities are only partially realized. The difficulties that will be faced by ASCI applications in the near future are foreshadowed by the challenges currently facing astrophysicists in making full use of the data they have collected over the years. For example, among other difficulties, astrophysicists have expressed concern that the sheer size of their data restricts them to looking at very small, narrow portions at any one time. This narrow focus has resulted in the loss of "serendipitous" discoveries which have been so vital to progress in the area in the past.

To solve this problem, a new generation of computational tools and techniques is needed to help automate the exploration and management of large scientific data. This whitepaper proposes applying and extending ideas from the area of data mining, in particular pattern recognition, to improve the way in which scientists interact with large, multi-dimensional, time-varying data.

Data-mining is a process concerned with uncovering patterns, associations, anomalies, and statistically significant structures and events in data. It is a multi-disciplinary area, incorporating techniques from statistics, machine learning, scientific computing, data management, and visualization. One of the key areas in data-mining is pattern recognition, namely, the discovery and characterization of patterns in image and other high-dimensional data.

Pattern recognition has a wide variety of applications to programs throughout the Laboratory, as explored below. A research-oriented team focused in this area would have strong programmatic influence. Furthermore, as large-scale data mining is currently in its infancy, any results related to scaling pattern recognition techniques to large-scale data would be of great interest to the research community in general.

2.1 Programmatic Interest

Pattern recognition and data-mining will be invaluable in helping sort through the large, complex data sets generated or collected by programs across the Laboratory. A sample of some projects that have already expressed strong interest in these ideas is indicated below. The following two examples, discussed in detail, are representative of several potential projects identified in multiple discussions held with Drs. Gary Carlson, Hank Shay, Morry Aufderheide (B Division), Charles Alcock (IGPP), and Dan Schikore (CASC).

2.1.1 Synthetic Radiographs

Understanding and modeling shock propagation is one of the tasks that is central to LLNL's stockpile stewardship mission. Radiographs are essentially a series of images similar to x-rays images. They are used to provide experimental data on detonation fronts passing through multiple materials. One of the critical steps in developing accurate models of this phenomenon requires the scientist to distinguish between multiple computational models based on comparison to the experimental data. This involves producing several series of synthetic radiographs from the computational models, then identifying which series is the best match to the experimental data.

The experimental images, however, are noisy representations of shock fronts passing through a variety of materials of different densities. They are difficult to interpret in a precise manner. Currently, comparisons between experimentally and computationally generated images are made based on the experience, or the "gut feeling" of the designer. It can be difficult or impossible for the scientist to quantify his or her gut feeling in precise mathematical terms.

The application and extension of pattern recognition techniques could be quite valuable in this domain. Expected contributions include:

1. Extension of the set of features being used to describe radiographs, potentially including pixel-based features, higher-level features that describe the shape and form of the shock

front, as well as features based on SVD and wavelets. Some of this work would involve clustering techniques to determine the shape and form of the shock front.

2. Expression of the experimental shock front as a set of fuzzy, or probabilistic, features to account for the noise in the image and the natural variability in its topography. We may be able to reduce the imprecision by measuring and eliminating certain types of sensor error (similar to band-pass filtering).
3. Automatically build a quantitative figure of merit that can be used to compare radiographs alongside of, or in place of, the scientist's gut feeling.

The construction of a figure of merit would provide a metric on the degree of radiograph similarity, as well as an improvement in the quality and repeatability (over time, and between scientists) of radiograph comparisons. Furthermore, a computable figure of merit can help speed overall model evaluation by automatically filtering out large sets of images that are not close enough to experimental data to warrant the scientist's attention. The end result of this exploration should be shock wave models that are more accurate, and scientists that have a deeper understanding of the physics involved.

The major computer science research focus here revolves around robust statistical methods for describing "fuzziness" in the experimental data, and relating that to the accuracy of the corresponding computational models. This research will be important in understanding how to measure the accuracy of pattern recognition techniques, and could play an important role in controlling the tradeoff between accuracy and effort.

2.1.2 Micro-lensing

Astrophysics is a program that contributes greatly to the public scientific legacy produced by LLNL. Its importance is underscored by its relevance to ASCI (note that one of the 5 ASCI alliance centers is on astrophysics).

Charles Alcock, head of the IGPP, has developed a world-renowned effort for detecting galactic dark matter in the form of MAssive Compact Halo Objects. Machos are detected via gravitational microlensing, an event that occurs when a Macho, in close alignment with a background star, acts as a gravitational lens, magnifying and distorting the stellar image. Microlensing refers to the situation in which the image distortion is not detectable, and the only visible effect is an apparent amplification. This increase in intensity is transient, lasting from 20-60 days.

The MACHO collaboration seeks to identify microlensed stars by taking two color CCD images of upto 10 million stars per night in the region of the sky that is of interest. This image data is processed in real time, generating reduced photometry data, and possible microlensed events are reported. The actual identification of microlensed events is done by a human after examining the light curves and the original images. The probability of observing such events is very low, for example, two years of photometry on 8.5 million stars in the Large Magellanic Cloud has revealed 8 candidate microlensed events.

The MACHO project currently has five terabytes of image data on tape, as well as 500 GB of photometry data that is on-line. The work done so far has targeted the reduced photometry data as it is of a manageable size. Unfortunately, the sheer volume of the image data has made any form of unassisted human analysis all but impossible. As a result, a large part of the data that has been gathered is still unexplored.

Large-scale pattern recognition techniques could be extremely beneficial in this domain. For example, the current sensitivity of micro-lensing detection, based on the photometry data, is low, implying that many such events are passing by unnoticed. Analysis of the feature-rich image data would be very useful in improving this sensitivity. Typically, when given a reasonable quantity of relevant data and a rich set of descriptive features, classification techniques from data mining and statistics have built models that are much more sensitive and accurate than humans. Applying these techniques to increase the current micro-lensing detection capacity is the first task.

Applying clustering techniques on the image data in a directed knowledge discovery mode is a second, highly attractive goal. It would help the astrophysicists find patterns that they have not thought of previously, for example, a correlation between spatial and temporal phenomena. These directions would represent the first steps towards an automated data exploration capability that should help add the element of serendipity back into the science of astrophysics.

From the computer science perspective, major research issues include scaling the clustering and classification algorithms to cope with the sheer size of the data available, pre-processing the data to remove various errors without losing useful information, and investigating ways of enabling interactive exploration of the data.

2.1.3 Other Areas

Many other projects and programs at LLNL could derive great benefit from large-scale data mining. Briefly, some of these are:

- **ASCI:** This domain is replete with examples of complex data analysis that is carried out through images. Some examples include pin dome and PINEX experimental data. This program will be generating large amounts of computational data, and currently has limited support for automated analysis of any form.
- **ASCI Advanced Visualization:** Pattern recognition techniques can be used as an aid to visualization, enabling improved navigation and discovery in large data sets, as discussed in a recent joint proposal for the establishment of an ASCI Advanced Visualization Technology Center. Locating and tracking features through transient data is an important part of this proposal. Work is also needed to make these techniques more interactive,

perhaps by using smaller representative samples of the data being visualized, or by using multi-resolution data.

- **Verification and Validation:** Computational models are evaluated by measuring how well they model the actual physical phenomena, and how sensitive they are to model errors, or variations in the input parameters. The predictiveness of the model is determined by comparing the model data to experimental data, when available. The work in representing and comparing to fuzzy classes, as described in the synthetic radiographs example, involves robust statistics, and stochastic modeling. This provides a good foundation for addressing verification and validation concerns.

- **Global Climate Change:** This domain has large amounts of data that must be analyzed for trends that occur over time. The modeling done here is very similar to the work described above.

- **Genome:** There are several interesting applications in this area as well. For example, spectrograph readings are an important source of information for determining protein function. These images again share characteristics very similar to the radiographs discussed above. Also, pattern recognition techniques can be used to determine the factors that influence 3-dimensional structures in proteins, as well as help predict these structures.

The problem of effectively exploring and characterizing large scientific data is not specific to any one domain. Indeed, as hinted at above, it is important to a wide range of programs supported by the Computation Directorate. Currently, there is no coordinated center of activity at LLNL that is looking for possible solutions to this problem. Furthermore, commercial data mining products do not scale to large, high dimensional data sets, and the research community is only beginning to address these issues - primarily in the context of business data.

3 Technical Issues

3.1 Overview of the Data Mining Process

The process of exploring and analyzing data, as formalized by the data-mining community, is an iterative multi-step process involving data preparation, search for patterns, knowledge evaluation, and refinement. This interactive process typically involves the following steps:

Data preparation: This includes understanding the application domain, possibly selecting and integrating a subset of the datasets, cleaning the data to remove noise or imput missing fields etc. This step is very dependent on the problem domain, and requires close interaction with the domain experts.

Search for patterns: This step includes feature extraction or finding important attributes to describe the objects of interest in the data, data and dimensional reduction to reduce the effective number of instances or variables under consideration, then selection and application of a data mining algorithm.

Knowledge evaluation: This includes interpreting the discovered patterns, visualization of the patterns, analysis of the accuracy of the resulting model, or evaluation of the degree of confidence one should have in the results, and finally possibly returning to one of the previous steps to refine the process.

The scalable pattern recognition work we propose in this project concentrates on the later half of the data mining process, namely, the search for patterns, and knowledge evaluation. Much of the effort in pattern recognition is shared across domains and has similar qualities independent of the domain. Suitably isolating the algorithms and research issues from the specific problem domains should allow a research effort to be applicable to many different applications.

For example, the data central to the synthetic radiographs (SRG) and the MACHO applications described above appear at first to be quite different. The SRG data is much smaller in aggregate than the MACHO data, the error in a CCD image has a different form than the error in a radiograph, and in MACHO, classification will be used to build models, while in SRG probabilistic models will be constructed directly out of experimental data. However, both of the domains are characterized by noisy 2D image data that represent 3D artifacts. The images are time series data, which has a substantial impact on the correlation of the data. The main goals in both domains are to extract features of interest in a pre-processing step, use those features to compare the images, and along the way improve the scalability of the techniques being applied.

The research issues outlined below focus on scalable pattern recognition, and therefore are targeted at the application independent aspects of these problems. In this way, the results of this research can be applied to various application areas around the Laboratory. Collaborations with the programs will supply the domain-specific knowledge that is needed, and assistance in applying the research results in domain-specific projects.

3.2 Research Issues

Data-mining, especially for large complex data-sets, is very much in its infancy. The algorithms with better predictive power are far too expensive to apply to anything but small and relatively low-dimensional data. The less complex approaches currently scale to medium-sized data. However, there are often difficult tradeoffs to make between the time needed for a data mining task, the expected accuracy of the results, the complexity and the power of the selected algorithm, the sample size of the data used, and the number of features used to build a pattern. One of the largest examples in a scientific domain in today's literature is of a sky mapping survey done at NASA JPL. The problem domain is similar to the MACHO project described above, except there is no time series data, and the total data size is about two orders of magnitude smaller. The JPL data set consisted of about 100GB of image data. From this data, 50 features were used to describe each object. The run time was measured in days.

There are strong programmatic reasons to scale pattern recognition techniques to data of the current size of the MACHO corpus, and the near future sizes of ASCI data. The following barriers, which all have a major impact on the overall viability and efficiency of a data mining algorithm, will have to be addressed in some way:

- **Size of data:** Manipulating such large data requires dealing with issues such as storage, and data transport across networks and through the storage hierarchy from tertiary storage to disk to main memory.

- **Type of data:** Handling data types that are native to individual domains requires mapping from native data formats into the flat data structures used by modern data mining algorithms. This mapping can explode the number of features of each object, and potentially lose information in the process.

- **Algorithmic Complexity and Efficiency of Implementation:** Data-mining algorithms are powerful learning devices, but they have high algorithmic complexity associated with them. For example, simple naive Bayes, a fast but limited predictor, can be linear in the number of features + training examples. Decision trees, which are excellent classification tools, can range from $O(nlogn)$ to $O(n^2)$ or worse, depending on the type of pruning done. Model induction algorithms (e.g. belief networks) with built-in assumptions that reduce complexity start at $O(n^4D^2)$ where n is related to the number of features, and D is the average domain size for the variables. More complex algorithms are exponential, making them infeasible for large data sets.

  Due to this high algorithmic complexity, these algorithms can be computationally time-consuming. Ideally, the process should be interactive, providing the end-user a quick way of analyzing their data. However, the implementation of current data-mining algorithms is often inefficient, and rarely uses the advances in high performance scientific computing to reduce compute time.

- **Quality of the results:** Data mining is frequently used as a decision making tool. As such, if the tool is poorly understood, then expensive and incorrect decisions could easily be made based on the output (e.g. treatment decisions in the medical world). Data mining tools generate models that have built in biases, and are often the product of heuristic, non-exhaustive searches through a problem space. Unfortunately the quality of these models (or inferences) are rarely measured, leaving the user to rely on intuition as to how far to trust the results.

  If data mining techniques are to be trusted tools of any profession where the (real and opportunity) cost of a mistake is high, then the quantification of the inference results must be addressed. However, there is little to no work currently being done on practical, formal, mathematical modeling for measuring the quality of the results generated by data-mining.

3.3 Summary of Approach

We believe that research with a focus on moving towards a more interactive model of computer-assisted data exploration and analysis has the highest potential benefit. To achieve this, we are considering:

- Techniques from high performance computing to improve the computational efficiency of the implementation of data-mining algorithms. This includes using the latest methods in numerical analysis, efficient implementation on a single processor, and use of parallel programming techniques and parallel processors.
- Hybridizing approaches from soft computing and computational science such as neural nets or genetic algorithms with singular value decomposition or wavelets. Hybrid algorithms typically produce much more accurate patterns in practice than the individual component algorithms.
- Applying and extending techniques from stochastic modeling and robust statistics in order to provide the information needed to evaluate and validate the computational models derived from the pattern recognition process. This information is crucial in making appropriate use of the results of the pattern recognition algorithms. It also plays a central role in allowing one to control the tradeoff between accuracy and the amount of effort expended to achieve that accuracy.
- Exploring the use of random sampling and multi-resolution in the tradeoff described above to enable interactive data exploration. This is useful in cases where the user is more interested in light exploration rather than deep analysis, or in domains where early detection of an event in progress can allow scientists to closely monitor the event.

4 Leveraging

There are significant resources that this effort will be able to leverage:

1. High performance computing competence - CASC and the Computation Directorate have a history of excellence in large-scale parallel computing, including parallel I/O and scientific visualization. Many of the issues faced in these areas are similar to the challenges facing large-scale data mining. The expertise is local, well within reach of this project.
2. Computing resources - We have access to a cluster of DEC 8400's, limited access to HPSS platforms, and the teraflop-capable ASCI machines.

3. Large-scale applications - LLNL houses efforts that are visible at the national level in stockpile stewardship, astrophysics, energy, materials science, environmental sciences, human genome, and more. All of these domains have similar characteristics regarding their need to analyze large, complex data.

In addition, we will be able to leverage expertise being developed in a funded LDRD project with the Human Genome effort: "Data warehousing and integration for scientific data management". One of the project goals is to initiate a data mining project in the genome area.

Finally, the CASC personnel involved in this proposal have strong research backgrounds in high performance computing, data mining and large-scale data management. Our team is well positioned to have significant programmatic impacts, as well as to influence both the theoretical and practical aspects of this new field of large-scale data mining.

5 Glossary

Data mining: An alternate name for Pattern recognition used by statisticians and database researchers. Data mining is one step in KDD (Knowledge Discovery in Databases), which is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. KDD includes the cleaning and pre-processing of the data, data-mining, and finally, the visualization and interpretation of the patterns.

Feature: A feature is any extractable measurement or attribute used in a pattern recognition system. Features can be low-level entities such as signal intensities, color, weight, pressure etc. or high level entities such as aspect ratio, Euler number etc.

Figure of merit: A metric that measures the goodness between experimental and computational models.

Neural nets: A more complex, predictive model than linear regression. Unfortunately, the resulting learned model is basically impossible for the domain scientist to understand, and model training times can be unpredictable and long.

Pattern: A pattern is essentially an arrangement or an ordering in which some organization of underlying structure can be said to exist. It can also be referred to as a quantitative or structural description of an object or some other item of interest.

Pattern recognition: Pattern recognition is the application of algorithms to extract patterns (models) in data. It deals with automatic techniques for partitioning or assigning input patterns or measurements into meaningful categories.

Statistics: A general method of reasoning from data. It is a basic approach shared by people in today's society to draw conclusions and make decisions in business and in life. It lets us communicate effectively about a wide range of topics from sales performance to quality of computational models to operational efficiency. Statistics is the way that we reason effectively about data and chance in everyday life.

SVD: Singular value decomposition is a technique used to decompose a matrix into several component matrices, exposing many interesting properties of the original matrix. The ability of the singular value decomposition to split a vector space into lower dimensional sub-spaces is used in pattern recognition to reduce the number of features under consideration. Also referred to as the Karhunen-Loeve transformation, principal component analysis, or the Hotelling transformation.

Technical Information Department • Lawrence Livermore National Laboratory
University of California • Livermore, California 94551