SANDIA REPORT

SAND2008-6109
Unlimited Release
Printed September 2008

Proceedings of the 2008 Sandia Workshop on Data Mining and Data Analysis

James M. Brandt, Daniel M. Dunlavy, Ann C. Gentile

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

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

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2008-6109
Unlimited Release
Printed September 2008

Proceedings of the 2008 Sandia Workshop on Data Mining and Data Analysis

James M. Brandt, Ann C. Gentile
Visualization and Scientific Computing Department
Sandia National Laboratories
P.O. Box 969, Mail Stop 9152
Livermore, CA 94551-9159
{brandt, gentile}@sandia.gov

Daniel M. Dunlavy
Computer Science and Informatics Department
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1318
dmdunla@sandia.gov

Abstract

In this document, we report the proceedings of the 2008 Sandia Workshop on Data Mining and Data Analysis. This year's workshop focused on the the data analysis capabilities and needs of the space systems, satellite, ground-based monitoring, and remote sensing communities. In addition to the extended abstracts of each presentation of the workshop, summaries of the discussion sessions and resultant recommendations of the workshop committee are given.

Acknowledgments

We would like to acknowledge the following groups who made this workshop possible:

Sponsor

- Sandia Intelligence Modeling and Assessment Program (SIMAP)

Program Committee

- Jim Brandt (Chair)
- Danny Dunlavy
- Ann Gentile
- Youssef Marzouk
- Jim Nelsen
- Mike Procopio

Review Committee

- Kim Denton-Hill
- Carol Harrison
- Suzanne Rountree
- David Womble

Contents

Appendix

Figures

Summary

In this document, we report the proceedings of the 2008 Sandia Workshop on Data Mining and Data Analysis. This year's workshop focused on the the data analysis capabilities and needs of the space systems, satellite, ground-based monitoring, and remote sensing communities. Presentations from the applications community provided perspectives on such capabilities and needs in the following areas:

- Nuclear Weapons Stockpile Transformation
- National Technical Means Imagery Analysis
- Remote Sensing Satellite System Design

Key issues presented in these talks included the need to handle real data in operational scenarios, areas of user pain, and robust tool requirements.

The following presentations from the data analysis community provided insight into capabilities, particularly through use-case scenarios:

- Bayesian Analysis for Classification of Chem-Bio Detector Data
- OVIS: Scalable, Real-time Statistical Analysis of Very Large Datasets
- The Cognitive Foundry
- Heterogeneous Ensemble Classifiers
- Situational Awareness at Internet Scale—Onset Detection of Extremely Rare Crisis Periods
- The PageRank Derby
- Multi-way Data Analysis and Applications
- Flexible Data Analysis and Visualization with Titan
- Physics Insights for Modeling and Data Analysis

Finally there was a wrap up session during which the data analysis community identified key concerns in the areas of community communication, technical and programmatic challenges, and funding.

This report includes summaries of the discussion sessions and extended abstracts of each of the presentations. The program committee further reports its recommendations on the following issues identified during the course of the workshop:

- Mechanisms to achieve better community communication
- Keeping common documentation current
- Funding – How to fund informatics community interaction and common documentation

1 Workshop Details

The intent of the workshop is to discuss research ideas, technical challenges, open questions, and potential for collaboration across departments, centers, and applications at Sandia in the areas of data mining and data analysis listed below. The workshop will comprise presentations by researchers from Sandia in these areas and group discussions, with the intent of understanding what we are doing today, how it fits into the world at large, and what seems promising to tackle next.

The themes of this year's workshop include the following:

Application Areas

- Space systems
- Satellite
- Ground Based Monitoring
- Remote Sensing

Data Science Area

- Data fusion
- High dimension pattern recognition
- Large-scale ensemble models
- Graph analysis of large data sets
- Clustering techniques
- Feature recognition
- Visualization
- Waveform analysis
- Image analysis
- Signal processing

1.1 Participants

Heidi Ammerlahn
Brett Bader
Justin Basilico
Travis Bauer
Zach Benz
Jon Berry
Dave Bodette
Jim Brandt
Pat Crossno
Christopher Davis
Bert Debusschere
Kim Denton-Hill
Karen Devine
Danny Dunlavy
Neal Fornaciari
David Gallegos
Terri Galpin
Ann Gentile
Damon Gerhardt
Antonio Gonzales
Eric Goodman
Carol Harrison
Bruce Hendrickson
Howard Hirano
Christopher Hogg
Mark Hollingsworth
Philip Kegelmeyer
Tammy Kolda
Kurt Larson
Jennifer Lewis
Shawn Martin
Youssef Marzouk
Jackson Mayo
Sue Medeiros
David Melgaard
Daniel Myers
Jim Nelsen
Daniel Pless
Clark Poore
Mike Procopio
Brandon Rohrer
Tim Shead
Michael Stickland
David Thompson
Derek Trumbo
David White

1.2 Schedule

8:00–8:15	Sign in	
8:15–8:30	Welcome and Introduction	
8:30–9:30	**Session 1**	**Moderator: Jim Brandt**
8:30–8:45	Nuclear Weapons Stockpile Transformation	Kurt Larson (5534)
8:45–9:00	National Technical Means (NTM) Imagery Analysis	Kurt Larson (5534)
9:00–9:15	Remote-Sensing Satellite Systems Design	Kurt Larson (5534)
9:15–9:30	Discussion	
9:30–9:40	Break	
9:40–10:45	**Session 2**	**Moderator: Ann Gentile**
9:40–10:05	Bayesian Analysis for Classification of Chem-Bio Detector Data	Bert Debusschere (8351)
10:05–10:30	OVIS: Scalable, Real-time Statistical Analysis of Very Large Datasets	Jim Brandt (8963)
10:30–10:45	Discussion	
10:45–10:55	Break	
10:55–12:30	**Session 3**	**Moderator: Mike Procopio**
10:55–11:20	The Cognitive Foundry	Justin Basilico (6341)
11:20–11:45	Heterogeneous Ensemble Classifiers	Danny Dunlavy (1415)
11:45–12:10	Situational Awareness at Internet Scale - Onset Detection of Extremely Rare Crisis Periods	Philip Kegelmeyer (8962)
12:10–12:30	Discussion	
12:30–1:30	Lunch (provided by workshop)	
1:30–3:00	**Session 4**	**Moderator: Danny Dunlavy**
1:30–1:55	The PageRank Derby	Karen Devine (1416)
1:55–2:20	Multi-way Data Analysis and Applications	Tammy Kolda (8962)
2:20–2:45	Flexible Data Analysis and Visualization with Titan	Tim Shead (1424)
2:45–3:00	Discussion	
3:00–3:30	Break	
3:30–4:30	**Closing Discussions**	**Moderator: Danny Dunlavy**
3:30–4:00	Recap of talks and discussions	
4:00–4:30	Steps forward	

2 Workshop Summary

2.1 Technical Presentations

2.1.1 Applications

This session consisted of interrelated presentations of three application areas, with particular regard to elucidating data analysis challenges and providing insight into the application user's needs and perspectives. In this section we highlight the major issues and recurring themes which arose during the presentations and subsequent discussions.

Nuclear Weapons Stockpile Transformation An open informatics problem in the area of Nuclear Weapons (NW) Stockpile Transformation is the analysis of sub-populations in the data. The main challenges associated with this problem include: (1) determining and distinguishing sub-populations in the data; (2) determining which sub-populations should be included/excluded in a given analysis; and (3) identifying the constituents and number of separable trends evinced by an overall data set. Note that the development of technical bases for trend analysis is *not* an open informatics problem, as this community and its collaborators have considerable expertise in this area already.

NW data used for analysis comes from a variety of sources, for example as output from the manufacturing and testing processes of systems and components. Potential data analysis needs include anomaly detection, methodologies and techniques for data fusion from various sources and of various data types, and determination of the uncertainty incurred in analysis given the relative confidences in the individual data sources.

It is not always obvious what data is of interest or how to rigourously define the data that is of interest. Of particular value are tool interfaces and capabilities which do not require the user to form or ask specific questions or to form specific hypotheses. Rather, users would like to select data of interest and have the tool obtain the related meta-data. If a user or tool detects a pattern in that data, that user would like to obtain via the tool if others have observed similiar patterns or trends in other (perhaps unrelated) data.

Those who review and assess the work of the analysts are technically knowledgeable. Analyses should include measures of confidence in the results and rigorous explanations of outliers and behaviors.

National Technical Means Imagery Analysis The problem space of electro-optic (EO) imagery analysis is change detection within local, well-defined target areas, where the normal state does not change often and is well established. In addition to image data, there may be supplementary inconsistent text data and numerical data.

The "big data" problem in this area does not center around issues of mining through big data, but rather around issues in data retention. Of interest are data analysis methodologies that can sift through live data feeds, or streams, in order to determine what data to retain for further analysis. Also of interest are techniques and methodologies for data reduction and compression.

Analysts' needs include increasing the confidence of their results and reducing the drudgery associated with analyzing huge amounts of transient data. Recommendations for researchers interested in working in this application area: (1) expect to invoke incremental change; (2) address established use cases first; (3) target capabilities that improve user efficiency and/or confidence; and (4) capitalize on Sandia's resources.

Remote Sensing Satellite System Design Technical solutions in the area of remote sensing satellite system design must balance operational constraints with the need for high quality analysis results. The problem regime involves variable spatio-temporal processing distributed across space and ground components. Distributing analysis in such a multi-tiered fashion requires consideration of data rates at each step in the overall configuration (sensor data rate >> payload data rate >> download data rate, and download data rate <<<< ground station data rate) as well as the different computational and storage capabilities at each tier.

Data analysis areas of interest include (1) techniques for feature characterization and change tracking; (2) false alarm detection and quantification of associated probabilities and confidence; (3) motion detection; and (4) compressive sensing. Drivers for data analysis in space applications include (a) low false positives; (b) confidence intervals and likelihood estimates associated with hypothesis testing; and (c) the time scales of responses for which the analysis is being used.

Experiences and Lessons Learned in User-focused Application Development In the course of developing user-focused applications, software solutions, tools, etc., the following have been identified as being **highly valued** by users working in the domain areas listed above:

- Methods solve *my* problem with *my* data in *my* context.
- Solutions and analyses are reliable and accurate.
- Software solutions are supported.
- Tools are easy to learn.
- Methods and tools can be easily incorporated into *my* workflow.
- Software supports security requirements (this may be key).
- Software is cost effective (possibly).

In contrast, the following are often **not important** to users in these domains:

- Algorithms
- Architectures
- Publications
- State of the art

Users often want active support, maintenance, advice, and guidance from developers with regards to using new software tools. Data scientists should understand that software development and support is not the same as algorithm development. The speaker estimates that less than ten percent of the work involved in delivering a solution to customers is algorithm development. The algorithm developer should stay engaged with the application developer even after initial development.

Discussion Topics

- *Automated and machine learning capabilities applied to satellite and image analysis applications.* There could be interest in active learning techniques to identify the best new information to collect/investigate. There are currently no automated methods for determining priority in choosing images or image areas for detailed analysis. This latter issue may be a long-term goal as the problem space does not address wide area image search, but rather small, well-defined areas of interest within areas being imaged.

- *Limitations of current algorithms given the operating/payload constraints.* The speaker believes that EO image analysis will require new algorithms. Note that the satellite applications community is a risk-adverse community that must consider carefully investing in and deploying new methods. The correct approach is to validate a capability fully on the ground, so that by the time it is installed in satellites it is a robust, well-tested solution.

- *Special circumstances of failure data (discussed by members of the audience, not the speaker).* Often failure data is binary data (e.g., success or failure) vs. continuous performance metric data. There is tremendous pressure for explaining why failure cases are not characteristic of the data as a whole. Challenges exist in the areas of data fusion and resolution (e.g., image resolution, time scales, measurement sensitivity) of the data being analyzed.

Session Summary Data analysis areas of interest in these domain areas include:

- Anomaly detection
- Change detection
- Data fusion
- Image analysis—particularly change in small, well-established areas

Data analysis capabilities must address real-world, not theoretical, problems:

- Choosing the correct sample or sub-population
- Working with noisy, uncertain, and/or obscure (image) data
- Quantifying confidence/uncertainty
- Determining what data to retain during the collection phase
- Automating tasks to eliminate drudgery

Data analysis methodologies and implementations are subject to constraints as well:

- Trade-offs in data rates, power, and processing capabilities at each step in the analysis process
- Requirement of low false positives
- Analysis duration must be tailored to acceptable response times
- Incomplete framing of user questions

Analysts want easy-to-use, reliable, supported tools, which include appropriate measures for addressing security requirements. Research and algorithms in and of themselves are not a selling point; in many cases appropriate algorithms to solve particular problems have been developed within a project group. Rather, it is tools that combine and present this information appropriately that are currently desirable. For new capabilities to be integrated into an application, active interaction with and support from the developer are required. The user community is most open to solutions that target their areas of pain. Initially, researchers and developers should address established use cases and plan on incremental change.

In determining aspects to address, keep in mind that as a lab, we are supposed to focus on differentiating technologies.

2.1.2 Data Analysis

There were three workshop sessions focused on algorithms and software tools for solving data analysis problems. Techinques were illustrated through use-cases in order to better enable the audience to extrapolate potential application of the techniques to their problem domains. In this section, we summarize the main points of the presentations and subsequent discussions.

Session 2

- *Bayesian Analysis for Classification of Chem-Bio Detector Data* (Appendix A.2)
- *OVIS: Scalable, Real-time Statistical Analysis of Very Large Datasets* (Appendix A.3)

The first talk addressed the use of Bayesian methods as a probabilistic classification framework. It highlighted the value of this technique for noisy data and for determining confidence measures in the resulting classifications. The methods were successfully applied to chemical sensor data to determine the probability of the presence of chemical agents.

OVIS[1] is an extensible framework for sensor data collection, statistical analysis, and data visualization. Its intent is to enable outlier detection on a reactable timescale via statistical spatial and temporal analyses. Practical considerations in the OVIS design include 1) the capability to process large and frequent data streams and 2) reliability of the framework and analysis methodologies in the event of system component failure.

The discussion during this session focused on the following issues and questions:

- How important are the factors of noise, interferents, and environmental effects on the analysis of chemical sensors?
- What is the impact of false positives, and is minimization of false positives the main driver for the analysis techniques?
- Processing and analyzing sensor data at the appropriate reactive timescales is crucial for success in these application areas.
- There is great need for probabilistic descriptions and confidence estimates in the general area of data analysis. The approaches presented here both address these needs.

[1] http://ovis.ca.sandia.gov

Session 3

- *The Cognitive Foundry* (Appendix A.4)
- *Heterogeneous Ensemble Classifiers* (Appendix A.5)
- *Situational Awareness at Internet Scale—Onset Detection of Extremely Rare Crisis Periods* (Appendix A.6)

This session focused on machine learning approaches to data analysis. Several machine learning software tools and projects were discussed, including the Cognitive Foundry[2], HEMLOCK[3], and AvatarTools[4].

The Cognitive Foundry is a unified collection of software tools for cognitive science and technology applications. The Foundry consists of the Cognitive Framework and machine learning algorithms. Automated knowledge capture is an important feature of the Foundry that is applicable to many data analysis problems, as domain-related data models can help subject matter experts more effectively analyze large sets of data.

HEMLOCK is a software tool for building ensembles of heterogeneous models and combining these models to improve overall performance in solving data classification problems. This approach of using heterogeneous ensemble classifiers to solve classification problems can alleviate the burden on data analysts of choosing optimal software parameters while also generally leading to improved performance.

AvatarTools is a set of classification algorithms based on decision trees. The presentation focused on cutting edge research associated with Hellinger trees, a new approach to solving classification problems that handles skew in the data classes better than other existing methods. This method has shown promise in several applications, and thus shows promise as a general classification technique.

The discussion during this session focused on the following issues and questions:

- How well do these methods scale?
- In using ensemble classifiers, how can we balance optimizing base classifiers versus optimizing the overall ensemble model?
- Can classifier models be trained on one set of data and used on another?
- How much data is needed to build accurate models?

These questions (and more) are currently being researched by the presenters of the talks in this session and related groups, and will hopefully be addressed directly as capabilities in future versions of the tools mentioned above.

[2] http://cognition.sandia.gov/Projects/CognitiveFoundry/
[3] http://software.sandia.gov/hemlock
[4] http://www.ca.sandia.gov/avatar/

Session 4

- *The PageRank Derby* (Appendix A.7)
- *Multi-way Data Analysis and Applications* (Appendix A.8)
- *Flexible Data Analysis and Visualization with Titan* (Appendix A.9)

The focus of the PageRank Derby presentation was the performance of a particular data analysis algorithm used in graph analysis, PageRank, as implemented across a variety of hardware architectures. This work illustrated the importance of choosing algorithms and hardware appropriate for a given task in order to maximize computing performance.

In the multi-way data analysis presentation, the use of multidimensional arrays (i.e., tensors) and associated algorithms for performing multiple factor analyses on large, sparse data was discussed. Examples of the methods applied to several data analysis problems illustrated the general applicability of this approach and the associated Matlab software, the Tensor Toolbox[5].

The final presentation on the Titan framework[6] for large-scale information visualization illustrated the framework's flexibility and extensilbility. The framework extends the popular visualization software library called VTK, which is used internally at Sandia (as well as by Sandia customers and a large external community as well). Titan is built around a parallel data and execution pipeline architecture, making it flexible for many data analysis and visualization tasks. Its modular plug-in capabilities also facilitate rapid prototyping of new algorithms and customizable front-end interfaces.

The discussion during this session focused on the following issues and questions:

- How important is load balancing of data in these applications? Load balancing is important, but more rigorous studies still need to be performed to assess the impact on particular applications of interest. The work on the PageRank algorithm at Sandia is currently performed using the Trilinos solver framework, which include interfaces to Sandia's load balancing software called Zoltan.
- How scalable are these methods? Can they run on laptops, or do we need high performance computing capabilities? For the Tensor Toolbox, the focus on handling sparse data allows for analysis of graph structures with millions of links. Additionally, Titan is built on top of VTK and ParaView, tools which run on workstations as well as the high performance computing resources at Sandia.
- For open source software development, is it possible to split off Sandia-specific solutions and/or algorithms from the open source code? This is currently addressed in Titan using two software repositories and keeping codes separate throughout the entire design process. Also, through plug-in capabilities, Titan can share code fragments and capabilities internally very easily.

[5] http://csmr.ca.sandia.gov/tgkolda/TensorToolbox
[6] http://vizrd.srn.sandia.gov/vizwiki/

2.2 General discussion during wrap up session

There was significant discussion around there being many "tool kits", both in existence and under development for various programs. There was some agreement that the community could benefit by better dissemination of information relating to this body of work. There were several suggestions made about using a repository or a SharePoint site for these tools, along with possible locations where it might reside. One suggestion was the Analyst Homepage (http://analyst.ran.sandia.gov), however these tools mostly target scientific computing applications in the area of engineering design and analysis. Another possible location was a "re-use" software repository (https://repository.sandia.gov/logiclibrary/login/CoreAssets). One major problem identified with regard to a repository effort is maintenance. It is easy to put up initial information but keeping it up to date is difficult and would require management support. Finally, the idea of creating a new resource specific to informatics resources was presented and discussed. On a related note, Travis Bauer (5634) has compiled some information on informatics efforts around Sandia in the form of quad charts for use by SIMAP in discussing Sandia'a informatics capabilities with potential customers.

2.2.1 Community-related issues

One of the big challenges for the informatics community at Sandia is how to maintain/foster interactions not only between people working in research areas, but also between researchers and potential consumers of their work. Some of the suggestions were as follows:

- Make the WDMDA workshop an annual event.
- Coordinate periodic informatics related brown bag seminars.
- Extend group meetings/discussions of Big Analysis of Data Community of Practice (BADCOP) to people interested in informatics across Sandia. Through group meetings and brown bag seminars, BADCOP explores problems inherent to the retrieval, storage, analysis and understanding of large amounts of diverse structured, semi-structured, and unstructured data.

Additionally, it was observed that what individual groups are trying to achieve is very compartmentalized/closed, whereas delivery of solutions can be broad. The solution to this is not obvious as a person is typically only invited into a group based on having relevant and necessary expertise; hence exposure of problems to a broader group that might collectively have solutions is currently a challenge.

2.2.2 Technical/programmatic challenges and issues

Two significant challenges discussed were how to foster the use of tools in the user community and how to facilitate collaboration within the tool development and research community. With respect to use of tools, training was identified as a key element. It was proposed that developers offer beginner level training courses on their tool kits to this audience and that they be willing to provide deeper training sessions to those interested. Within the developer community there were two solutions proposed to fostering more collaboration. One was that a funding source be created specifically to develop strategic partnerships between toolkit developers, and the other was the

formation of a "toolkit council" which Zach Benz (6343) volunteered to spearhead. The council agenda would be to meet every few months for updates and planning.

2.2.3 Communication of work at SNL

Management is looking to staff for guidance on solutions, and staff should take the initiative on things such as forming councils, setting up wikis, etc., as they see utility in such efforts. Managers at the workshop encouraged staff along these lines and urged them to not be daunted by the possibility of failure, as even a short lived effort could have long term positive effects.

2.2.4 Funding issues

The areas identified that have need of funding are as follows:

1. Documentation of the proceedings of this workshop
2. Repository for informatics tools, and
3. Development of strategic partnerships between toolkit developers.

- SIMAP will fund the program committee this year to:
  - Publish a SAND report on the workshop
  - Put together a set of questions and issues that came up in the workshop
  - Stand up a wiki to put up results of the workshop
- BADCOP group could sponsor web site repository.
- No funding source was identified for development of strategic partnerships between toolkit developers.

2.2.5 Additional Problem Characteristics of Interest

Throughout the workshop, many technical problems were presented and discussed; during the general discussion, several additional technical problems and challenges were discussed as well. Staff from the application areas are interested in solutions for handling fuzzy data, subjective labels on data, data fusion, and seamless storage and sharing of data. Some of these issues have unique characteristics or challenges in particular application domains, but in general, they are of importance to anyone attempting to analyze data.

3 Recommendations

We make recommendations for two different time regimes. Short term recommendations are made on issues that we believe can be significantly addressed over the next year. Additionally we propose that longer term issues, some of which we identify here, be addressed by the formation of a working group of volunteers from this community.

Short term issues and recommendations:

- **Mechanisms to achieve better community communication:**
  - It became clear during this workshop that not only was there a need for closer communication between application developers and the informatics research community but also within the Sandia informatics research community itself. In many cases better communication can help ease the strain on limited resources by identifying areas of duplication as well as lack. Resources currently dedicated to duplicative efforts could be deployed to cover areas in which we are currently lacking while new and closer collaborations could also serve to propel duplicative efforts to faster solution.

    * **Recommendations:**
      · Set up a wiki that all attendees will have accounts on and can contribute to.
      · Identify people within each center/department to act as a point of contact that can attend discussions, seminars, and such.
      · Hold regular (every couple of months) one hour sessions for presenting work and needs (at an appropriate level).
      · Encourage application groups to hold brown bag/seminar sessions at a level of detail that allows researchers to attend and get a sense of problems.
      · Researchers should start attending existing brown bag sessions within application groups as allowed.
      · Based on understanding of needs and areas of interest, individual researchers should pursue more in depth understanding of application groups needs.

- **Keeping common documentation current:**
  - There was a lot of discussion around the problem of keeping documentation about a changing communities efforts current. While putting initial documentation into a repository is a relatively easy task the difficult problem is maintenance. Making a single person responsible to periodically collect data on changes and modify a repository doesn't scale well and would require specific funding.

    * **Recommendations:**
      · Put up a wiki with read/write access for internal staff only.
      · Have each person putting content on the wiki be responsible for keeping their content current.
      · Have an automated reminder quarterly or semi-annually about reviewing/updating content sent to everyone who has contributed content.

- **Funding**
  - In much of the current funding and political structure, it is far easier to get funding for theoretical research, prototype development, and specific analysis tasks, than it is to get support for building the infrastructure necessary to turn data analysis techniques into useful tools.
  - Financial support explicitly for fostering better community communications is non-existent.
    * **Recommendations:**
      · Take advantage of funding from organizations such as SIMAP and BADCOP where their mission overlaps with this communities needs.
      · Continue to fund workshops and interactions out of informatics projects.
      · For the short term people should either charge relevant projects (ones whose features are being documented) or donate time to keep their content up to date.

Longer term recommendation:

Our recommendation for longer term issues is the formation of a working group comprised of both researchers and application developers to work out longer term issues. All activities associated with longer term issues, including identification of the issues, plans for solutions to these issues, and implementation of the resulting solutions will **require funding** that does not currently exist for such activities. We recommend that a source and level of funding be identified to support a working group to address issues such as the following:

- Roles and goals of Sandia's informatics capabilities both internal and external to Sandia
  - Balance between research and application development
  - Balance between prototyping and hardened tools
  - If a goal is to translate the informatics capabilities into useful tools (either general or program focused), how can that be accomplished programmatically?
  - If a goal is to translate the informatics capabilities into useful tools, what additional technical requirements should we be investigating in developing analyses (e.g., performance vs. footprint tradeoffs)
- How to minimize duplication of effort while encouraging exploration of capabilities and implementations. This is a difficult issue as many efforts look similar on the surface while underlying characteristics may be very different. For example, different implementations of a similar technique might be addressing speed, data size, compute or storage constraints, post-processing or real-time data feeds with different temporal characteristics.
- Installation, training, and support
- Determinion of mechanism(s) for communicating results of the working group to groups with a vested interest in the area of informatics research (e.g., LDRD Office, DOE Office of Science, centers working in this area) in order to maximize impact of the working group.

4 Workshop Feedback

4.1 Questions

A post-workshop survey was sent out to participants. From the 46 participants there were 15 survey responses. The survey questions are listed below, and were aimed at eliciting feedback about the workshop, collect ideas on how to best move forward in the areas of algorithm development and application of informatics capabilities, and survey how much interest there is at Sandia in the are of informatics.

1. Talks were targeted to my area of application, research, interest, etc.

2. The technical level of interest was approriate for audience.

3. There was enough time for discussion.

4. There was enough chance to contribute my point of view during discussions.

5. There was a good mix of centers/division representation at the workshop.

6. I would be interested in attending next year's workshop.

7. I would be interested in helping plan next year's workshop.

8. I would be interested in participating in an ongoing lab-wide working group in the area of data mining, data analysis, informatics, etc.

9. I would be interested in contributing to an online resource of seminar announcements, software development, projects, and general lab-wide information in the area of data mining, data analysis, informatics, etc.

10. I would be interested in using such an online resource in my own work.

11. I would be interested in being a point of contact for my department, group, center, or division in lab-wide efforts in building relationships in the area of data mining, data analysis, informatics, etc.

12. I would like to be placed on a mailing list dedicated to discussions of lab-wide efforts in the areas of data mining, data analysis, informatics, etc.

4.2 Results

Item	Strongly Agree	Agree	No Opinion	Disagree	Strongly Disagree
1	3	8	4	0	0
2	7	6	2	0	0
3	5	6	4	0	0
4	5	6	4	0	0
5	6	3	5	1	0
6	11	3	1	0	0
7	1	5	7	2	0
8	6	8	1	0	0
9	6	7	2	0	0
10	6	7	2	0	0
11	5	4	4	2	0
12	11	4	0	0	0

Comments from Participants

- The technical level of Kurt's talk was about right but in many places important details were omitted so that in the end an algorithm-type person such as myself could not be helpful without further discussion. I expect everyone in the audience was in the same boat, in which case – why have the talk?
- It was a good workshop and great to get everyone together that is interested in these areas from across the labs. One suggestion would be to have at least part of the workshop (or a separate workshop) without requiring a Q clearance level (hopefully including uncleared), to make sure that we can include everyone who is interested in Data Mining.
- I was pleased to hear the first talk, on the needs of users (despite its being abridged due to sensitivity); I hope future workshops, and the discussion area, continue that. We are all working hard on the technology, and being reminded of the need for that technology (and probably the need for creative ways to combine and leverage our various technologies and areas of expertise) is very useful.
- Great workshop! Thank-you so much for providing this leadership for Sandia!
- A lot of the research type talks dealt with specific code implementations/libraries. That is all well and good to hear about. But it would be nice to work in Matlab. I would be interested in porting a few functions into Matlab to play with. Maybe something like this would be a good project for an online-collaborative effort. There seemed to be a good spectrum of topics covered from "What are the program needs", "What are folks tinkering with" to "How are we using this for real applications". Thanks to whoever put the program together.
- I appreciated the diversity of topics presented - from a needs point of view to a solutions point of view.

A Workshop Abstracts

A.1 Perspectives on User Needs for Data Exploitation

Author: Kurt Larson (5534)

A principal challenge in developing production applications for data exploitation is for the Subject-Matter Expert to identify the practices of a user community that are both painful and solvable via novel data exploitation. I believe that pain is unfortunately a necessary motivator for any user community to admit that it is powerless over its problem and to be open to help from outside its inner circle. Figuring out the methods whereby user pain can in fact be solved via data exploitation is a necessary part of the marketing and negotiation process. I will discuss perspectives on this problem drawn from three domains: Nuclear Weapons Stockpile Transformation, National Technical Means (NTM) imagery analysis, and remote-sensing satellite systems design.

Nuclear Weapons Stockpile Transformation

Nuclear weapon systems are now being managed as "unlimited lifetime" machines, a paradigm shift from traditional "design life" perspectives. Monitoring programs are being implemented to detect changes in performance due to aging, backed up by formal, quantitative trend analysis. A principal difficulty in evaluating a trend is determining whether the data includes outliers or sub-populations that should be excluded. Outlier uncertainty causes engineers to wonder about the integrity of trend analysis, yet identifying "causal" explanation of outliers can require significant effort. An opportunity exists to discover data mining and exploitation techniques that bridge between trend analysis and explanatory evidence for outliers or sub-populations. To be successful, the capability will need to accept imprecise descriptions of the nature of the outlier, and be compatible with a wide variety of legacy and current data sources including those at other NWC sites.

A key perspective on this problem is that developing the technical basis for trend analysis in NNSA's "Quantification of Margins and Uncertainty" (QMU) program, is NOT the problem that affords an opportunity to break into the community using data exploitation. The nuclear weapons community includes many PhD statisticians working formally as Reliability or Independent Assessments engineers, as well as others with undergraduate and post-graduate statistics in formal roles such as System, Component, or Surveillance engineers. They are quite happy to debate amongst themselves and believe, probably correctly, that their expertise is sufficient to solve the "trend analysis" problem. Technical capability for trend analysis is not an area of customer pain.

The NW community anticipated, and has realized in many examples, cases in which there is significant uncertainty over which data should be included in a population for trend analysis. Sometimes this uncertainty is realized before the trend is even attempted based on known information; other times the uncertainty arises when a population that was expected to be comparable seems to have non-uniform behavior. In the latter case, the engineer evaluating the trend has a high-risk decision: should the suspect data be included or excluded? Choosing incorrectly may lead to incorrect decisions in stockpile management. Finding evidence to support the decision is ad-hoc and

could lead nowhere. The domain of search includes many formal data sources, for example: ISIS, Record-of-Assembly, Image Management System (IMS), Significant Findings Investigation (SFI), WETL Test Archives (lab tests), Flight Test Archives, Kansas Citys PPTD Database, and other component-specific product production and testing databases. Further complicating the search, these databases exist in multiple computing environments and need-to-know groups and present different interfaces to the users.

A data exploitation opportunity lies in linking the tools in which unexpected trend anomalies are likely to be discovered, such as Sandias Wave application or MiniTab, with "Google-like" document search results providing a likelihood that a document contains information relevant to the cause of the anomaly. A 2008 pilot project led by Brett Bader is investigating a linkage between Wave and an informatics-based exploitation of WETL TAR archives. While it is hoped that this pilot will demonstrate high value in data exploitation in this area, it barely scratches the surface of the data types that are available to be integrated, and does not include resources to build the capability into production tools. System, Surveillance, and Expert User Services engineers recognize that progress in this area could be highly beneficial and have been supportive of the pilot project.

National Technical Means Imagery Analysis

Visible-band National Technical Means imagery (EO) provides the human analyst with the most intuitive medium of all remote sensing technologies, and it remains a mainstay of intelligence collection and analysis. Due to its complexities, analysis of EO is an intensely manual task. Analyst workloads are carefully managed. In general, the capability to collect images exceeds the capability to utilize them. Only the most important sites can be monitored routinely, and few retrospective analyses of historical imagery can be conducted. In general, with important exceptions, practitioners have been over-promised, under-delivered by AI technologies in the past and are thus skeptical of new proposals. The user community is served by a billion-dollar-scale enterprise architecture. This enterprise is built and serviced by industry giants who are protective of their status. Despite these obstacles, opportunities and sponsors exist to investigate methods to improve the utilization of EO systems.

The sources of pain in this system are several. Although analysts may think that EO capabilities are under-utilized, this pain is abstract and does not affect them on a daily basis. There are several concrete areas of pain. The analyst wonders whether looking at additional imagery would improve their assessments. Another "pain" is drudgery: analysts probably go into their jobs hoping to watch dynamic situations where every image adds to the story. Instead, many spend a considerable amount of time assessing the state of vital, yet dull and unchanging sites.

I believe data exploitation opportunities exist in helping the analyst monitor imagery feeds and archives to identify potentially relevant information, and in figuring out how to monitor relatively unchanging situations in a highly automated manner. The technical complexity of EO imagery makes success very difficult to achieve, and baby steps are appropriate.

Given the institutional-enterprise barriers to entry at the large scale, in my view the successful

initiatives will first be harmoniously complementary to existing capabilities in non-threatening ways. Later, once trust is established, opportunities to merge capabilities into existing or next-generation systems may emerge.

Remote Sensing Satellite System Design

A remote sensing satellite system necessarily separates the data-rich environment of the sensor from the resource-rich environment of its ground station. The downlink bandwidth is seldom sufficient to transmit data at the rate of the sensor. For a satellite operating under the paradigm that a single mission has priority over all others, the tension between data availability and processing power is accommodated by up-front mission engineering across many constraints to derive a relatively static downlink management policy providing variable performance across missions. Where multiple missions have similar priority, however, it is impossible to devise a single policy that promises optimal performance across all mission requirements. The generally risk-averse approach to variable on-board data processing is yielding to the necessity of achieving sensor performance specifically tailored to the current mission (or simultaneously-operating missions). An opportunity exists for innovative injection of new types of data exploitation within the satellite payload that achieves maximal performance in any mission. Any solution must also satisfy total-system constraints including sensor physics, payload processing power, downlink capacity, automated and human-in-the-loop ground station processing, command and control, probabilities of detection and false alarm, and event characterization requirements.

An obvious approach is to move time-tested ground-station algorithms into embedded systems in the satellite payload, yet this suffers from an inflexibility if the ground station algorithms are highly sequential or branching in nature. A more interesting approach, and one that is more likely to satisfy multi-mission operability, is to create area-of-interest processing policies on the payload in which the spatio-temporal sampling of data varies across the field of view. AOIs could be established to monitor for different things in different areas of the field-of-view, or to monitor the same areas of the field of view in different ways; output of the AOIs would be passed through the downlink management policy. The inclusion of on-board buffering to enable retroactive extraction of features at high fidelity is another option in system design. Compressive sampling and other techniques driven by modern information theory are also interesting possibilities for downlink data management.

Two other areas of interest for satellite systems data exploitation will be mentioned.

A.2 Bayesian Analysis for Classification of Chem-Bio Detector Data

Authors: Bert Debusschere (8351), Habib Najm (8351), Jose Ortega[7]
For more information: {bjdebus,hnnajm}@sandia.gov

Sandia has developed extensive expertise in various sensor platforms for the detection of CBRNE (Chemical, Biological, Radiological, Nuclear and Explosive) agents. The determination of whether an agent is present requires a classification of the sensor data stream based on attributes that are characteristic of specific agents. This paper outlines the use of Bayesian methods as a probabilistic classification framework that is ideally suited to handle noisy data and to assess the confidence in the detection accuracy. The method is illustrated for the analysis of μChemLab$^{\text{TM}}$viral signatures and of stochastic nanopore sensing device data. Applications of this work extend to a wide range of detector platforms.

In the Bayesian context, the classification of agents from characteristic attributes in their detector signals is formulated probabilistically. The posterior probability of the unknown analyte being a specific agent is proportional to the likelihood of observing the given data attributes for this specific agent, multiplied by the probability, based on prior information, of this agent being present. To classify an unknown analyte, the ratios of the posterior probability of it being a specific agent to the probability of it being each of the other agents in the training set are computed. The logs of these ratios, referred to as Bayes factors, provide the probabilistic inference, based on the data and prior information, of the identity of the unknown analyte. This probabilistic framework is well suited to handle noisy data, and the magnitude of the Bayes factor provides a measure of confidence in the classification.

In the application to μChemLab$^{\text{TM}}$, viruses are identified based on the characteristics of their protein electropherogram, obtained by capillary gel electrophoresis after lysing[1]. The key data attributes that allow discrimination between different viruses are the relative distances between characteristic peaks in the electropherogram, often corresponding to high-copy-number proteins. For various viruses (MS2, RSV, Vaccinia, EBV, T2 and T4), sets of key peaks were identified, acting as the signature of each agent. The likelihood functions for observing these attributes were then determined from training data for all agents under consideration. When applied to non-training data, the accuracy of the classifier was 66/69 or 95%. The three failures were easily explained by problems during the electrophoretic analysis (e.g. low sample concentration or injection problems) or peak detection by the preprocessing tool. The failures were of the detection type, where the classification algorithm detected "other". There were no false alarms (detecting a virus when none is present). The lowest positive Bayes factor observed was about 6.5, hence a ratio of probabilities among the (correctly) detected agent and the next potentially detected agent of about $e^{6.5}$ or $\sim$700x; clearly a high-confidence answer. The average of all observed positive Bayes factors was $\sim$25.

In stochastic sensing, nanopores (e.g. protein channels formed by the bacterial toxin α-hemolysin) in membranes are used as sensitive and selective detectors for metal ions and organic macromolecules such as proteins or DNA. The passage of the individual molecules or ions through the

[7]Formerly SNL

pore causes transient blockages in the ionic current when a voltage is applied across the membrane. Attributes of the current signature (amplitude, duration of shut/open times, and frequency of blockages) can be used to identify the nature of the analytes and to estimate their concentrations. The Bayesian classification approach was applied to the detection of binary Zn-Co mixtures using simulated α-hemolysin data, and of polymers using experimental α-hemolysin data. As attributes, the observed transitions from the upper and lower current level to any other level and the duration of the current gaps both at the bottom and the top were used. The performance of the classification in terms of accuracy, sensitivity, and robustness was investigated as a function of the noise in the data and the amount of samples taken and characterized using detector Receiver Operating Characteristic (ROC) curves. The methodology was found to perform very well, even for high-noise, sparse data sets, including detection of binary mixtures.

Overall, the use of Bayesian methods for agent classification in the μChemLab$^{\text{TM}}$and stochastic sensing applications was found to be both computationally efficient and highly effective, delivering reliable classification results with quantified confidence.

References

[1] Fruetel, J. A., et al., *Rapid identification of viruses using microfluidic protein profiling and Bayesian classification*, Anal. Chem., submitted 2008.

A.3 OVIS: Scalable, Real-time Statistical Analysis of Very Large Datasets

Authors: Jim Brandt (8963), Bert Debusschere (8351), Ann Gentile (8963), Jackson Mayo (8963), Philippe Pebay (8963), David Thompson (8963), Matthew Wong (8963)
For more information: http://ovis.ca.sandia.gov, ovis-help@sandia.gov

OVIS (http://ovis.ca.sandia.gov) is an extensible framework for sensor data collection and statistical analysis. Its original development targeted early detection of anomalous component behavior in large scale computational platforms. Leveraging the fact that such platforms are composed of large aggregations of "identical" hardware components, OVIS's analytical methodologies involve statistical spatial and temporal analyses for detecting outlier behaviors. Here, we describe OVIS's extensible scalable framework, consisting of data collection, storage, visualization, and analysis components and present use cases, both mature and under investigation.

Scalable Framework

Large scale computational platforms are composed of aggregations of tens of thousands of "identical" hardware components, each capable of reporting on the order of 100 variable values per second ($\sim$10 million values/second). Thus, we required a collection and storage mechanism capable of scaling several orders of magnitude past this current scenario to accommodate platforms expected to emerge over the next decade. Data collection and storage are both distributed for scalability and are distributed to different components of the system for robustness. In order to accommodate storage and quick retrieval for analysis we use a distributed database model.

Visual Analysis

Because OVIS is intended for use when systems are deployed (where debugging epistemic startup issues is the primary concern) as well as for continued operation (where predicting and detecting aleatory failure events is paramount), visualization tools must be available so that humans can develop mental and then mathematical models to test. Because the project's primary use case is monitoring large scale computational platforms and because we have found that the environment plays a large role in the system's state, the visualization tools in OVIS are focused on rendering 3-D views of the components. (Components' values are represented by color in this display.) However, system configurations are easily modified so that other types of components can be rendered and our work on the SNIFFER project, where networks of chemical sensors are deployed in various venues, has been able to make use of the visualization tools to illustrate sensor data. The visualization tool kit (VTK) on which OVIS is built also allows us to rapidly prototype new visualizations and we are working to leverage plotting tools being developed by the TITAN project for use in OVIS.

Numerical Analysis

OVIS provides ways to "learn" descriptive statistics (minimum, maximum, mean, variance, skewness, and kurtosis of a metric), correlative statistics (mean, variance, covariance, linear correlation

coefficient, and linear regression lines calculated on pairs of metrics), and bivariate Bayesian parametric models (fits a metric of a normal, log-normal, or exponential distribution whose parameter(s) is/are function(s) of another metric). Contingency statistics (joint PDF, conditional probabilities, information entropies) have been recently added to VTK and will soon be incorporated into OVIS.

The output of a "learn" analysis is a model, which may be utilized in itself, or may further either be taken as ground truth to find outliers in a given dataset ("monitor") or, conversely, be tested against a dataset ("validate"). In either case, the dataset may be the same dataset used in computing the model, or a different one. Such models can also come from any other origin such as manufacturer specification, expert knowledge, or simply be a desired range of operation. The "validate" mode of execution is to inform the user as to the applicability of a model to a particular dataset. It can be utilized, by repeatedly running "validate" analyses using the same model, but with successive datasets, to assess the progressive drift of baseline behavior with respect to an initial model. Such a drift may be entirely normal and even expected or indicate that something is deteriorating. These 3 modes of execution address the following cases: (1) There is no existing model, or an existing model has been found to not fit current data and so a new model is being calculated. (2) A window of time has passed and a model is being validated against the incoming data. (3) Incoming data is being compared for classification against a valid model. Note that several of these use cases rely on the tasks above being performed at regular intervals.

Analysis and Use Cases

1. *High Performance Computing (HPC).* We have used OVIS in the HPC environment in several ways:

   (a) Visual consideration is a quick-and-easy, non-computationally intensive way to consider the cluster as a comparative ensemble, rather than as singleton components. Visual analysis has enabled diagnosis of the root cause of thermal gradients, hot spots, and temperature conditions exceeding the recommended operating regime on several of our platforms to be associated with particular air flow issues. Additionally, visual discovery of global geographical patterns, such as variation of CPU temperature with height or location, have allowed us to distinguish environmental effects from true abnormal behavior. In an example case, visual consideration led us to fit the temperature distribution with a quadratic model, with statistical outliers then determined relative to this model. This is an effective technique for systems in a non-uniform environment where the environmental effects may mask the effects we are trying to discern.

   (b) Our statistics tools have allowed discovery of components with properties that deviate significantly from what we expect given the properties of a large population of similar components. In the HPC regime, this has been used for discovery of faulty fan controllers and voltage regulators, mis-calibrated analog to digital converters, and a variety of anomalous conditions whose root causes we have yet to discover.

   (c) By meaningfully quantifying the "health" of components and detecting significant changes in this indicator with time, we seek to dynamically place important or long running tasks on more reliable components and less critical tasks on less reliable components. This strategically increases the overall reliability of the system where it counts. Under our assumption that abnormality can be an indicator of less reliability, we define numerical representations of "health" as a function of variations in the evinced single component

statistical moments from those of the group. We use relative ranking of this value as an indicator of the relative health of the components. We then look for temporal change in this value to determine significant state change. This work is extensible to resource allocation in other regimes. We are currently extending this work to the security regime where the health value would be a quantification of the likelihood of a component being significantly compromised.

(d) We have ongoing development work in Bayesian time series analysis with the goal of learning and classifying characteristics of normal network traffic both within HPC high speed interconnect networks and in the enterprise. The goal in both areas is to be able to identify outlier behaviors. The reasons, however, are quite different. In the HPC interconnect, anomalous behavior can be an indication of either software or hardware fault conditions. In the enterprise anomalous behavior can be an indication of malicious behavior on the part of an either an intruder or an insider. This type of work would be extensible to areas where one wants to detect anomalous characteristics in a long lived data stream.

2. *Application Performance Optimization Tools.* OVIS is collaborating with Sandia's ASC performance tools project to develop advanced, scalable data analysis capabilities for software optimization on HPC systems. The goal is to enable simple runtime feedback on the degree and cause of application code inefficiencies, without requiring manual interpretation of performance counters. Using profiles for a set of simple programs called kernels, we use regression models to estimate the amount of wasted runtime in a given region of an application and also identify the kernels most similar to its profile. Thus developers can focus on the most important regions to optimize, and obtain insight from the known inefficiencies exhibited by the similar kernels. A preliminary implementation for single-processor applications has shown the possibility of accurate diagnosis of a real-world performance issue. This analysis approach illustrates the idea of exploiting measured variables (e.g., the available performance counters), whose individual meanings may be obscure, and transforming them to a superposition of previously tabulated reference cases (e.g., meaningful performance patterns). The method is useful when the link between an observation and its explanation is indirectly available through (possibly non-orthogonal) variables that collectively contain significant information. Each variable should be measured in a consistent way, but its meaning need not be understood.

3. *Chemical Sensors.* In various applications, the detection and proper characterization of events of interest requires analysis of data streams containing a large number of variables. One example is the detection of chem-bio attacks using a sensor network, where the data stream to be analyzed consists of signals from many different sensors, each sensor tuned to respond to specific substances. In practice, however, these systems tend to be plagued by false alarms due to the fact that these sensors commonly also register ordinary substances besides the harmful ones they are intended to detect, the low-level presence of toxic substances in harmless products (e.g. household cleaners), and general background noise. Also, in many cases, one substance will trigger multiple sensors, introducing correlations between different dimensions in the data stream. In order to properly separate true events from false positives in such cases, a capability is being developed in OVIS for a high-dimensional multi-variate probabilistic characterization of the "normal" state of a sensor network. This will help to reduce false positives as "true" events will be detected as outliers from the normal state. This approach is being applied to the Sandia SNIFFER project.

A.4 The Cognitive Foundry

Authors: Justin Basilico (6341), Kevin R. Dixon (6341), Zachary Benz (6343)
For more information:
Justin Basilico: jdbasil@sandia.gov
on the SRN: http://cognition.sandia.gov/Projects/CognitiveFoundry

Introduction

The Cognitive Foundry is a unified collection of tools for Cognitive Science and Technology (CS&T) applications. We have designed the Foundry to be a robust, extensible, and reusable software platform to support the development of intelligent agent models, particularly those automatically created from data. The Foundry has two primary components designed to facilitate construction: the Cognitive Framework and Machine Learning packages. The Cognitive Framework provides design patterns and default implementations of an architecture for evaluating theories of cognition, as well as a suite of tools to assist in the building and analysis of theories of cognition. The Machine Learning package provides tools for populating components of the Cognitive Framework from domain-relevant data using machine-learning and data-mining techniques. We have designed the Foundry to be a robust, extensible platform to support research, rapid prototyping, and system development by providing reusable software components and algorithms designed to support a wide variety of data-driven development needs.

The design of the Cognitive Foundry has followed a graduated interface approach. That is, the Cognitive Foundry is built on top of a set of well defined, hierarchical interfaces. The Cognitive Foundry then provides one or more default implementations of these interfaces. However, developers can always create their own tailor-made implementations if existing ones do not meet their needs, allowing researchers to test new ideas and hypotheses quickly. Since tools in the Cognitive Foundry provide functionality at the interface level, new components can automatically utilize functionality provided by existing components in the Foundry by conforming to a defined interface. There are several benefits to this interface-centric component- based approach. It provides an easy mechanism for customizing existing object implementations in the Foundry. It also gives the ability to pick the specific objects from the Foundry that are useful for a certain application. Finally, it creates an integration point for many applications, which defines an easy transition path from research to deployment.

Cognitive Framework Package

The Foundry's Cognitive Framework is a modular software architecture for cognitive simulation. The Cognitive Framework itself is a collection of interfaces, which allows Framework users to either leverage the existing tools in the Framework or specify different implementations to fit their specific needs in order to test new ideas and hypotheses. The Cognitive Framework is designed so that different, and possibly competing, elements of a "theory of cognition" can be instantiated as desired. This is accomplished by having a Cognitive Module perform some aspect of a psychologically plausible cognitive process. A Cognitive Model, then, contains a collection of Cognitive Modules whose purpose is to instantiate some aspect of cognition. Modules pass information to one another

through cognitive elements (or "cogxels") in the model state. The model also provides state encapsulation containing the sufficient information needed to allow a model to resume execution later, or on another machine, without altering the results of a simulation.

There are currently two implementations of the Cognitive Framework: a lightweight implementation and a concurrent implementation. The lightweight implementation, also known as the Cognitive Framework Lite, is specifically designed for being embedded in high-performance simulations with many (usually small) cognitive models running in a single process. Thus, it contains compact data structures, a fast update loop, and the ability for modules to share parameters to conserve memory. The concurrent implementation is designed to run on multi-core/processor computers to allow separate Cognitive Modules within a Cognitive Model to execute in parallel. Thus, it can support running large models in a single process by making use of all available resources.

Machine Learning Package

The Cognitive Foundry's Machine Learning Package provides a wide variety of optimized, verified, and validated general- and special-purpose algorithms for machine learning and data mining: the analysis and characterization of large datasets, function minimization, parameter estimation, prediction, and categorization. It is also designed to support using automated knowledge-capture techniques to populate cognitive models. The package is highly extensible, meant for allowing the rapid-prototyping of applications based on machine learning and the development of new or experimental algorithms and architectures. Thus, it provides a toolbox for both researching new machine learning or data mining techniques and also tailoring techniques to new applications.

Typically, in machine learning, there are various conflated components, such as:

- the object being created
- the learning algorithm used to create the object
- the data upon which the algorithm operates
- the domain-specific information for the algorithm
- the performance measure
- the statistical validation routine

The Machine Learning package separates each of these components with well- defined interfaces. This allows users of new functions to use existing learning algorithms and, conversely, creators of new learning algorithms to test their ideas on different functions. This allows rapid prototyping and experimental testing of different algorithms, approaches, and function approximators and categorizers. The package accomplishes this through the systematic use of interfaces and generics to encapsulate the needs of each algorithm, including their inputs, outputs, and parameterizations. We followed an object-oriented design for the entire package so that the different algorithms utilize common, interchangeable subcomponents, such as cost functions and statistical validation. This approach greatly simplifies the integration of exiting machine-learning algorithms to new problems and, conversely, to apply new machine-learning algorithms to existing problems and datasets.

A.5 Heterogeneous Ensemble Classifiers

Authors: Danny Dunlavy (1415), Sean Gilpin (1415)

Introduction

Recent results in solving classification problems indicate that the use of ensembles classifier models often leads to improved performance over using single classifier models [1, 2, 3, 4]. In this talk, we discuss heterogeneous ensemble classifier models, where the member classifier models are not of the same model type. A discussion of the issues associated with creating such classifiers along with a brief description of the new HEterogeneous Machine Learning Open Classification Kit (HEMLOCK) will be presented. Results for a problem of text classification and several standard multi-class test problems illustrate the performance of heterogeneous ensemble classifiers.

Heterogeneous Ensemble Classifiers

Classification is the task of learning a target function that maps data instances to one of several predefined categories. These target functions are also called classifiers, classifier models, and hypotheses. We refer to a classifier constructed or learned from an ensemble of different types of classifiers is as a *heterogeneous ensemble classifier*. Note that such classifier models are also referred to as hybrid ensemble classifiers.

There are several challenges associated with learning heterogeneous ensemble classifiers. First, the choice of which base classifiers (i.e., ensemble member classifier models) needs to be determined. Performance of classifiers differs across different data sets, and thus choosing the collection of classifiers that will best classify a given set of data is often a difficult task. Each base classifier can be parametrized in many different ways, and thus an understanding of how these parameters are correlated within each base classifier as well as across the ensemble is key to classifying data sets accurately.

Fusion and Selection

A further challenge is combining base classifiers effectively, so that the performance of the ensemble classifier is better than that of the individual classifiers. There are two basic strategies for combining classifiers in an ensemble: fusion and selection [5]. Ensembles that use selection try to find the best classifier ensemble member that is most capable of correctly classifying a particular instance. Ensembles that use selection are also known as cooperative ensembles. In contrast to selection, fusion methods make use of the outputs of all of the classifiers to try determine the label of an instance. Voting is an example of fusion: each of the classifiers in the ensemble is given one vote and all of the votes are counted towards deciding which output label should be chosen. Ensembles that use fusion are commonly referred to as competitive ensembles. There are three levels at which classifiers output can be combined using fusion: label, ranking, measurement. At the label level the ensemble will only use the one class label that each of the base classifiers determines is correct. For ranking, base classifiers in the ensemble provide a ranked list of class labels reflecting how likely

each class is marked as the correct label for each data instance. Finally, at the measurement level each of the base classifiers provides output that is intrinsic to the particular learning algorithm used. Typically, measurements consist of probability distributions of the class assignment for each instance.

Diversity

It has been shown that the strength of an ensemble is related to the performance of the base classifiers and the lack of correlation between them (i.e., model diversity) [3, 4]. One way to decrease the correlations between the classifiers while increasing or maintaining the overall performance of the ensemble classifier is to include base classifiers derived from different learning algorithms such as decision trees, neural networks, perceptrons, support vector machine, etc.

HEMLOCK

HEMLOCK is a new software tool for constructing, evaluating, and applying heterogeneous ensemble data models for use in solving classification problems involving data with continuous or discrete features. HEMLOCK consists of various data readers, machine learning algorithms, model combination and comparison routines, evaluation methods for model performance testing, and interfaces to external, state-of-the-art machine learning software libraries.HEMLOCK uses XML for all input and output, and standard readers and writers are being used for data input and output. Data models are created by a variety of supervised learning methods: decision tree and random forest inducers plus a linear perceptron learner as part of HEMLOCK along with interfaces to the methods available in the WEKA software library of machine learning algorithms. Evaluation methods for assessing individual model performance include accuracy computation, confusion matrix generation, receiver operating characteristics (ROC) analysis, and area under the curve (AUC) analysis. Methods for combining heterogeneous models into a single ensemble model include majority voting and parameter regression.

Applications

In this workshop, two applications of heterogeneous ensemble classification will be discussed: e-mail classification [2] and image classification. For the e-mail classification problem, a heterogeneous ensemble of random forest, naive Bayes, and perceptron classifiers were used as base classifiers (both as individual classifiers and in homogeneous ensembles). The image classification problem consisted of labeling the number represented in handwritten digits, and heterogeneous ensembles were created using the new HEMLOCK software framework.

References

[1] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, *A comparison of decision tree ensemble creation techniques*, IEEE Trans. Pat. Recog. Mach. Int., 29 (2007), pp. 173–180.

[2] J. D. Basilico, D. M. Dunlavy, S. J. Verzi, T. L. Bauer, and W. Shaneyfelt, *Yucca mountain LSN archive assistant*, Technical report, SAND2008-1622, Sandia National Laboratories, 2008.

[3] S. Bian and W. Wang, *On diversity and accuracy of homogeneous and heterogeneous ensembles*, Intl. J. Hybrid Intel. Sys., 4 (2007), pp. 103–128.

[4] W. Wang, D. Partridge, and J. Etherington, *Hybrid ensembles and coincident failure diversity*, in Proc. International Joint Conference on Neural Networks, 2001.

[5] K. Woods, K. Bowyer, and W. P. Kegelmeyer, *Combination of multiple classifiers using local accuracy estimates*, IEEE Trans. Pat. Recog. Mach. Int., 19 (1997), pp. 405–410.

A.6 Situational Awareness at Internet Scale—Onset Detection of Extremely Rare Crisis Periods

Authors: Philip Kegelmeyer (8962), David Cieslak (8962)
For more information:
David Cieslak: dcieslak@cse.nd.edu, http://www.nd.edu/~dcieslak/
Philip Kegelmeyer: wpk@sandia.gov, csmr.ca.sandia.gov/~wpk

It would valuable to have early warning of disruptions to the Internet. One aspect of Internet function is the Border Gateway Protocol (BGP), which handles the delivery of packets between various autonomous domains. In particular, autonomous routers use BGP communications for "announcements" and "withdrawals", that is, to announce the detected availability or unavailability of various Internet routes.

Certain statistics of these announcements and withdrawals might be useful in distinguishing between normal and abnormal operation of the Internet[3]. Practical use of these statistics in real-time monitoring is stymied, however, by the need for robust pattern recognition methods that can operate under conditions of extreme skew. In other words, the disruptions, though critical when they occur, nonetheless are occurring only a tiny fraction of the time: in the last decade of Internet activity, there has been roughly twenty cumulative days of "disruption", an incidence rate of 0.5%. This problem of imbalanced data afflicts many applications, and is becoming even more prevalent as data volumes grow.

"Bagged" ensembles of decision trees have been shown to be a simple, robust and accurate method for the detection of events in noisy data[1]. But even bagged trees falter in the face of such extreme skew. After all, if a disruption event shows up only 0.5% of the time, then a classifier could generate the trivial rule of classifying everything as "normal" and still be 99.5% accurate.

Thus we require both analysis methods *and* accuracy metrics more appropriate to skew data. Hellinger trees are a new decision tree analysis method that have been shown to be statistically significantly more accurate and robust than Infogain trees[2] (the best current decision tree method) when applied to skew data. Further, Hellinger trees have also been shown to be statistically significantly superior to pre-processing skew data with SMOTE, which was previously the best known corrective for skew data.

These advantages have been established, however, only in the context of the use of a *single* tree, not ensembles. Accordingly, we are investigating the application of ensembles of Hellinger trees to skew data, starting with the integration of Hellinger trees into the "AvatarTools"[4] tool set. AvatarTools is a Sandia developed and maintained suite of executables for supervised machine learning via ensembles of decision trees. For the current purposes, its differentiating capabilities are the ability to automatically choose the proper ensemble size, and the implementation of a slew of skew correction methods, including SMOTE. The latter serve as a comparative baseline for Hellinger trees.

In experimentation with AvatarTools over a wide variety of test data, we have been able to show that ensembles of Hellinger trees are statistically significantly superior to ensembles of Infogain trees on skew data, and also significantly statistically no worse than Infogain trees on balanced data. In sum, this means that anyone using decision trees should always use Hellinger trees, as they never

hurt and often substantially help.

Accordingly, we are investigating the use of ensembles of Hellinger trees in the analysis of BGP data for the detection of disruptions in Internet activity. We have acquired real world BGP data[5] that captures activity before, during, and after a variety of known historical disruptions[8], including worms, fiber cuts, distributed denial of service attacks, and power blackouts. We have used this real data to construct and study skew scenarios as extreme as 1000:1, and present one disruption scenario in which bagged ensembles of Hellinger trees do outperform bagged ensembles of Infogain trees, with the performance gap widens monotonically with increasing skew.

References

[1] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, *A comparison of decision tree ensemble creation techniques*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (2007), pp. 173–180.

[2] D. A. Cieslak and N. V. Chawla, *Learning decision trees for unbalanced data*, in Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Antwerp, Belgium, September 2008.

[3] D. Dou, J. Li, H. Qin, S. Kim, and S. Zhong, *Understanding and utilizing the hierarchy of abnormal BGP events*, in Proceedings of the 2007 SIAM International Conference on Data Mining, 2007, pp. 467–472.

[4] W. P. Kegelmeyer, K. Buch, and D. Cieslak, *AvatarTools*. www.ca.sandia.gov/avatar.

[5] RIPE NCC, *RIPE routing information service raw data*. www.ripe.net.

[8] We thank Max Planck of the New Mexico Institute of Mining and Technology for his invaluable work in acquiring, pre-processing, and truthing this data.

A.7 The PageRank Derby

Authors: Karen Devine (1416), Jonathan Berry (1416), Steve Plimpton (1416)

Introduction

Massively multithreaded parallel architectures such as Cray's MTA and XMT are proving to be highly effective for graph analysis algorithms. By providing uniform memory access times for data with much irregularity and little locality, these machines have demonstrated excellent scalability for a wide range of graph-based algorithms. However, no apples-to-apples comparisons between these architectures and our traditional distributed memory architectures had been performed using realistic input data.

We make the first such comparisons using Google's PageRank method as the algorithmic kernel and synthetic datasets with power-law vertex degree distributions. One might expect PageRank to favor traditional distributed-memory architectures, as PageRank reduces to matrix-vector multiplication with floating point arithmetic. However, the choice of data clearly favors the multithreaded architectures. By comparing PageRank performance using both distributed memory and massively multithreaded paradigms, we seek to assess the effect of algorithm and data on architecture choice.

PageRank

PageRank [5] computes the importance of web pages based on the importance of pages that link to them. The importance of page s increases if s is pointed to by other important pages. The share of importance that s receives from page t is inversely proportional to the number of pages that t links to.

PageRank models the web as a directed graph $G(V, E)$, with each vertex $v \in V$ representing a web page and each edge $e_{ij} \in E$ representing a hyperlink from v_i to v_j. The probability of moving from v_i to another vertex v_j is $\alpha/d_{out}(v_i) + (1-\alpha)/|V|$, where α is a user-defined parameter (usually 0.8-0.9), $d_{out}(v)$ is the outdegree of vertex v, and $|V|$ is the cardinality of V. The first term represents the probability of following a given link on page v_i; the second represents the probability of moving to a random page. For pages with no outlinks, the first term is $\alpha/|V|$, indicating equal likelihood to move to any other page.

MultiThreaded Graph Library Implementation

In the MTGL [1] implementation of PageRank, rank propagation is accomplished through adjacency list traversal. Our results were obtained with an underlying compressed sparse row data structure, but the same code would run on other graph representations. A key requirement for scaling is that the code must be written so that a single thread spawns the loop that processes all in-neighbors of a given vertex. This enables the compiler to generate hotspot-free code.

Distributed-Memory Implementation

The distributed-memory implementations of PageRank represent the graph as a matrix A [4], with matrix entries $A_{ij} = \alpha/d_{out}(v_i)$ if vertex v_i links to v_j. The PageRank algorithm, then, is simply a power-method iteration in which the dominating computation is matrix-vector multiplication $Ax = y$, where x is the PageRank vector from the previous iteration. Terms representing random links could conceptually be included in A, but they are more efficiently handled as adjustments to y. Rows or non-zeros of A are uniquely assigned to processors, along with the associated entries of the PageRank vector x. Interprocessor communication is needed to gather x values for matrix-vector multiplication and to sum partial products into the y vector. Most communication is point-to-point communication, but some global communication is needed for computing residuals and norms of x and y.

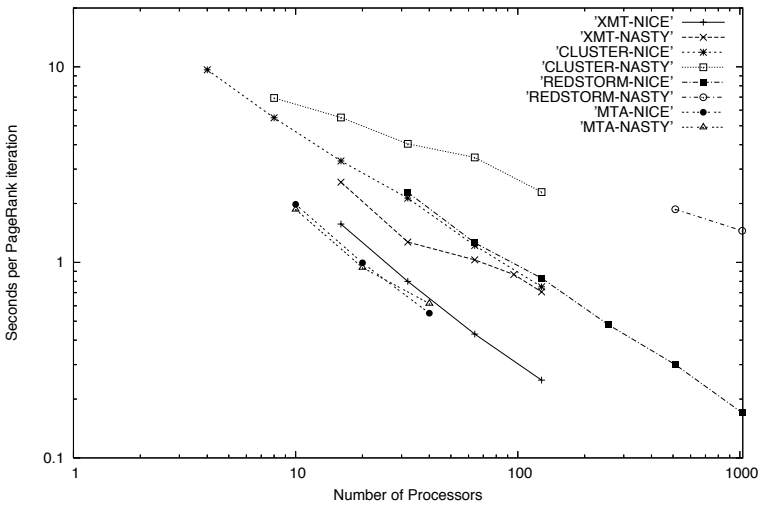


Figure A.1. Preliminary PageRank Derby results.

Experimental Data

Our experimental data are R-MAT graphs [2]. R-MAT graphs are recursively generated graphs with power-law degree distributions. They are commonly used to represent web and social networks. They are generated using only four parameters a, b, c, and d, which represent the probability of an edge being generated in one of four recursive quadrants of a matrix representing the graph. We used two different R-MAT data sets, each with average vertex degree of eight and 25 R-MAT levels (i.e., 2^{25} vertices). The "nice" data set uses R-MAT parameters $a = 0.45$, $b = 0.15$, $c = 0.15$ and $d = 0.25$; the resulting maximum vertex degree is 1108. The "nasty" data set uses R-MAT parameters $a = 0.57$, $b = 0.19$, $c = 0.19$, and $d = 0.05$; the resulting maximum vertex degree is $230,207$.

Results

We ran our experiments on Cray's XMT and MTA as well as Sandia's RedStorm and a small cluster called Odin. Cray MTA and XMT both support a global address space, hiding memory latency by using 128 instruction streams per processor. The MTA has 220 MHz processors and a modified Cayley network; the XMT has 500 MHz processors and a 3D-Torus network. RedStorm is a distributed memory parallel supercomputer with two Dual-Core AMD 64-bit 2.4 GHz Opteron processors per node, 2GB memory per node, and 9.6 GB/s link bandwidth. Odin has two AMD Opteron 2.2GHz processors and 4 GB of RAM per node; nodes are connected with a Myrinet network. For each data set, we computed page rank on each architecture; we present the time for one pagerank iteration in Figure A.1. For the nice data set, all implementations scaled well, with the XMT and MTA taking less time per iteration than the distributed memory machines. With the nasty data set, load imbalance limits the scalability of the distributed memory implementations; the maximum number of non-zeros of A assigned to a processor is an order of magnitude greater than the average. Load-balancing techniques like those in Zoltan [3] have the potential to improve the distributed memory performance. The XMT also struggles somewhat with the nasty data set. Previous experiments on the MTA-2 led us to expect little or no change in the scaling plots. The hitch in the XMT's nasty data plot was an unwelcome surprise that has yet to be explained. However, we note that the XMT's strong scaling on these data from 64 to 128 processors resumes a near-optimal slope. Parallelizing compiler artifacts may eventually explain the hitch.

Conclusions

Our results are preliminary. We have yet to run on exactly the same instances, and there are artifacts of each implementation that need further exploration. However, several themes are emerging.

- **Distributed memory clusters *can* process large, unstructured datasets in certain contexts.** Our simple algorithm applied to nasty data found strong scalability through at least 1000 processors.

- **Massively multithreaded architectures *can* outperform microprocessor-based clusters, even with floating point-intensive algorithms on datasets that are only mildly unbalanced**. The MTA/XMT machines have been spoken of only as boutique, pointer chasing vehicles. Before this study, there was little reason to expect 500 MHz processors without floating point units to compete with modern microprocessors in any floating point context. We explore sparse matrix multiplication on mildly unbalanced data, however, and see this surprising result in our nice data plots.

- **Realistic informatics data require programmer intervention for distributed memory architectures, but not for massively multithreaded architectures** The slopes of the strong scaling plots suggest that without load balancing, 10,000 Red Storm processors might not match 32 XMT processors on realistic informatics data. The effectiveness of load balancing remains to be measured.

References

[1] J. W. Berry, B. Hendrickson, S. Kahan, and P. Konecny, *Software and algorithms for graph queries on multithreaded architectures*, in Proceedings of the 21st International Parallel and Distributed Processing Symposium, March 2007.

[2] D. Chakrabarti, Y. Zhan, and C. Faloutsos, *R-mat: A recursive model for graph mining*, in Proceedings of the Fourth SIAM International Conference on Data Mining, April 2004.

[3] K. Devine, E. Boman, R. Heaphy, B. Hendrickson, and C. Vaughan, *Zoltan data management services for parallel dynamic applications*, Comp.in Science and Engineering, 4 (2002), pp. 90–97.

[4] A. N. Langville and C. D. Meyer, *A survey of eigenvector methods for web information retrieval*, SIAM Review, 47 (2005), pp. 135–161.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd., *The pagerank citation ranking: Bringing order to the web*, Technical report, Stanford Digital Library Technologies Project, 1998.

A.8 Multi-way Data Analysis and Applications

Authors: Tamara G. Kolda (8962), Brett W. Bader (1415)

The following abstract comprises excerpts from Kolda and Bader [9].

Introduction

We consider the analysis of multidimensional or N-way data, stored as *tensors*. A tensor is called higher-order if it has three or more modes. A third-order tensor has three indices as shown in Figure A.2. Decompositions of higher-order tensors have applications in psychometrics, chemometrics, signal processing, numerical linear algebra, computer vision, numerical analysis, data mining, neuroscience, graph analysis, and elsewhere. Two particular decompositions can be considered to be higher-order extensions of the matrix singular value decomposition (SVD): CANDECOMP/PARAFAC (CP) decomposes a higher-order array as a sum of rank-one tensors, and the Tucker decomposition is a higher-order form of principal component analysis. Sandia has developed the MATLAB Tensor Toolbox for working with higher-order arrays. We (quickly) survey multi-way decompositions, applications (with an emphasis on signal processing and image applications), and the MATLAB Tensor Toolbox developed here at Sandia.

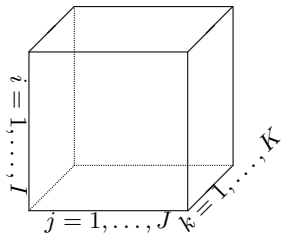


Figure A.2. A third-order tensor: $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$

Notation

The *order* of a tensor is the number of dimensions, also known as ways or modes. Vectors (tensors of order one) are denoted by boldface lowercase letters, e.g., $\mathbf{a}$. Matrices (tensors of order two) are denoted by boldface capital letters, e.g., $\mathbf{A}$. Higher-order tensors (order three or higher) are denoted by boldface Euler script letters, e.g., $\mathcal{X}$. Scalars are denoted by lowercase letters, e.g., a. The ith entry of a vector $\mathbf{a}$ is denoted by a_i, and element (i, j, k) of a third-order tensor $\mathcal{X}$ is denoted by x_{ijk}. The jth column of a matrix $\mathbf{A}$ is denoted as $\mathbf{a}_j$.

PARAFAC/CANDECOMP

The CP decomposition [5, 8] factorizes a tensor into a sum of component rank-one tensors. For example, given a third-order tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, we wish to write it as

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r, \tag{1}$$

where R is a positive integer, and $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$, and $\mathbf{c}_r \in \mathbb{R}^K$, for $r = 1, \ldots, R$. The *factor matrices* refer to the combination of the vectors from the rank-one components, i.e., $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_R \end{bmatrix}$ and likewise for $\mathbf{B}$ and $\mathbf{C}$. Elementwise, (1) is written as

$$x_{ijk} \approx \sum_{r=1}^{R} a_{ir}\, b_{jr}\, c_{kr}, \text{ for } i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K.$$

This is illustrated in Figure A.3. Methods for computing CP are surveyed in [9].

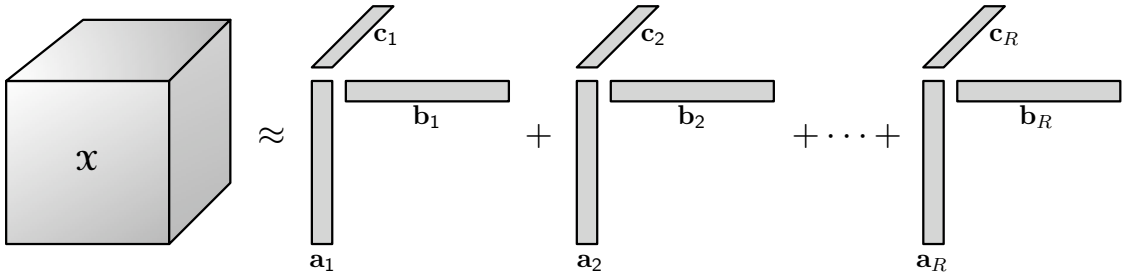


Figure A.3. CP decomposition of a three-way array.

There are many applications of CP described in [9]; here we mention a few that may be relevant. Sidiropoulos, Bro, and Giannakis [10] considered the application of CP to sensor array processing, and there are other applications to telecommunications listed in [9]. CP also has important applications in independent component analysis (ICA); see [6] and references therein. Bauckhage [4] extends discriminant analysis to higher-order data (color images, in this case) for classification.

Tucker

The Tucker decomposition is a form of higher-order principal component analysis. It decomposes a tensor into a core tensor multiplied (or transformed) by a matrix along each mode. Thus, in the three-way case where $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, we have

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr}\ \mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r \tag{2}$$

Here, $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$ are the factor matrices (which are usually orthogonal) and can be thought of as the principal components in each mode. The symbol $\times_n$ denotes tensor-times-matrix multiplication in mode-n; see [9] for details. The tensor $\mathcal{G} \in \mathbb{R}^{P \times Q \times R}$ is called the *core*

tensor and its entries show the level of interaction between the different components. Elementwise, the Tucker decomposition in (2) is

$$x_{ijk} \approx \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr}\, a_{ip}\, b_{jq}\, c_{kr}, \quad \text{for} \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K.$$

Here P, Q, and R are the number of components (i.e., columns) in the factor matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, respectively. If P, Q, R are smaller than I, J, K, the core tensor $\mathcal{G}$ can be thought of as a compressed version of $\mathcal{X}$. In some cases, the storage for the decomposed version of the tensor can be significantly smaller than for the original tensor; see Bader and Kolda [2]. The Tucker decomposition is illustrated in Figure A.4.

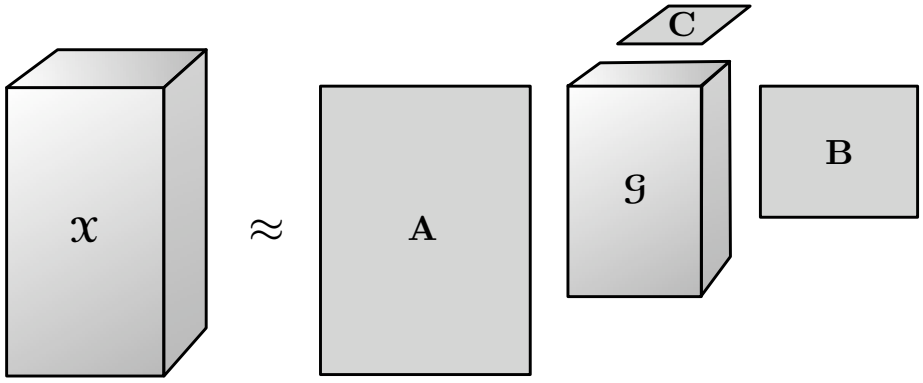


Figure A.4. Tucker decomposition of a three-way array

There are many applications of Tucker decomposition described in [9]; here we mention a few that might be of interest. De Lathauwer and Vandewalle [7] consider applications of the Tucker decomposition to signal processing. Vasilescu and Terzopoulos [11] pioneered the use of Tucker decompositions in computer vision with TensorFaces and have since extended the work to recognition, human motion, and more. Vlassic et al. [12] use Tucker to transfer facial expressions.

MATLAB Tensor Toolbox

The MATLAB Tensor Toolbox, by Bader and Kolda [1, 2, 3], is a general purpose set of classes that extends MATLAB's core capabilities to support operations such as tensor multiplication and matricization. It comes with ALS-based algorithms for CP and Tucker, but the goal is to enable users to easily develop their own algorithms. The Tensor Toolbox is unique in its support for *sparse* tensors, which it stores in coordinate format. The Tensor Toolbox also supports structured tensors so that it can store and manipulate, e.g., a CP representation of a large-scale sparse tensor. The Tensor Toolbox is freely available for research and evaluation purposes.

Summary

The CP and Tucker decompositions are useful for multiway data analysis, particularly in image and signal processing. There are many other decompositions as well that may also be of use (see

[9] for many more details). Sandia has developed the MATLAB Tensor Toolbox [1, 2, 3] which is a useful package for developing software for working with multiway data.

References

[1] B. W. Bader and T. G. Kolda, *Algorithm 862: MATLAB tensor classes for fast algorithm prototyping*, ACM Transactions on Mathematical Software, 32 (2006), pp. 635–653.

[2] ——, *Efficient MATLAB computations with sparse and factored tensors*, SIAM Journal on Scientific Computing, 30 (2007), pp. 205–231.

[3] ——, *MATLAB tensor toolbox, version 2.2.* http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/, January 2007.

[4] C. Bauckhage, *Robust tensor classifiers for color object recognition*, in Image Analysis and Recognition, vol. 4633 of Lecture Notes in Computer Science, Springer, 2007, pp. 352–363.

[5] J. D. Carroll and J. J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition*, Psychometrika, 35 (1970), pp. 283–319.

[6] L. De Lathauwer and J. Castaing, *Blind identification of underdetermined mixtures by simultaneous matrix diagonalization.* To appear in *IEEE Transactions on Signal Processing.*, 2007.

[7] L. De Lathauwer and J. Vandewalle, *Dimensionality reduction in higher-order signal processing and rank-$(R_1, R_2, \ldots, R_N)$ reduction in multilinear algebra*, Linear Algebra and its Applications, 391 (2004), pp. 31 – 55.

[8] R. A. Harshman, *Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis*, UCLA working papers in phonetics, 16 (1970), pp. 1–84. Available at http://publish.uwo.ca/~harshman/wpppfac0.pdf.

[9] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Review. to appear (accepted June 2008).

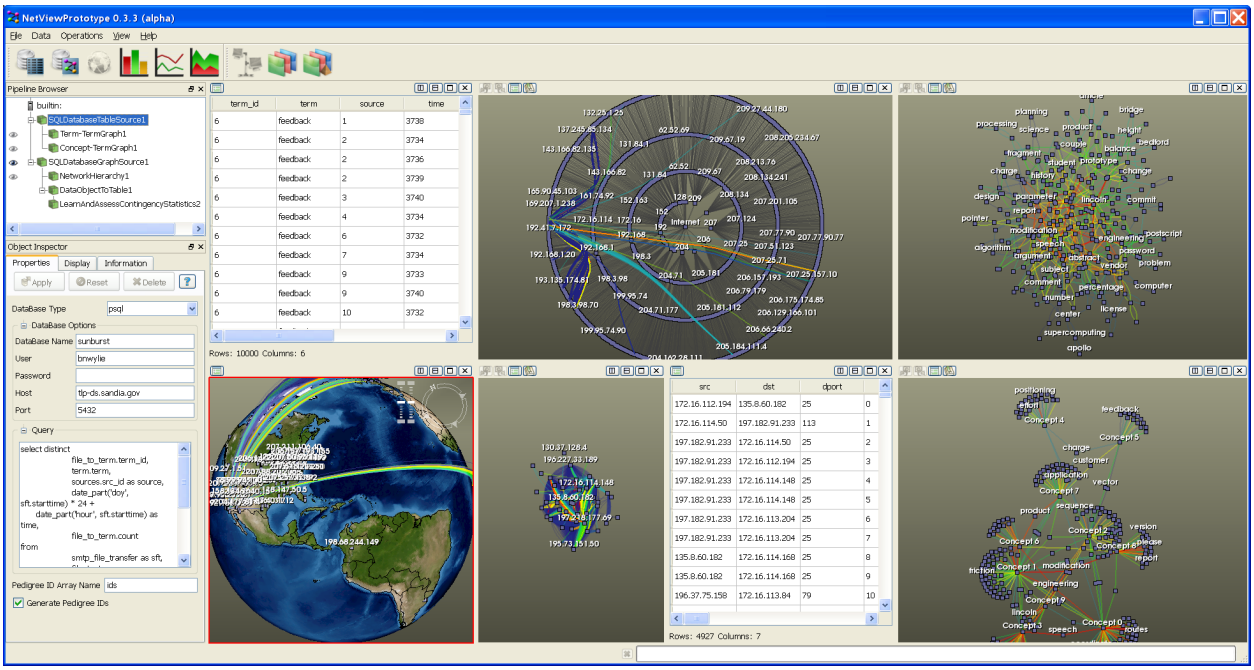
[10] N. Sidiropoulos, R. Bro, and G. Giannakis, *Parallel factor analysis in sensor array processing*, IEEE Transactions on Signal Processing, 48 (2000), pp. 2377–2388.

[11] M. A. O. Vasilescu and D. Terzopoulos, *Multilinear analysis of image ensembles: TensorFaces*, in ECCV 2002: Proceedings of the 7th European Conference on Computer Vision, vol. 2350 of Lecture Notes in Computer Science, Springer, 2002, pp. 447–460.

[12] D. Vlasic, M. Brand, H. Pfister, and J. Popović, *Face transfer with multilinear models*, ACM Transactions on Graphics, 24 (2005), pp. 426–433.

A.9 Flexible Data Analysis and Visualization with Titan

Author: Timothy M. Shead (1424)



Background

The technical challenges of informatics are similar to those of large-scale scientific visualization – terabyte datasets encompassing millions of entities and billions of relationships will require HPC clusters, desktop delivery, and reasonable response times for effective analysis. Using tools such as VTK and ParaView, SNL has processed, analyzed, and visualized some of the largest and most complex simulations in existence. The Titan Informatics Toolkit expands these tools to support ingestion, processing, analysis, and display capabilities for informatics data. Titan provides a flexible architecture for the integration and deployment of algorithms relating to intelligence, semantic graphs, and information analysis.

Architecture

Titan is based on a "filters and pipes" architecture, where individual components are data-sources, data-filters, or data-sinks that are interconnected to form data-processing pipelines. Decomposing problems into the source-filter-sink model captures high-level processes in an intuitive way and encourages exploration through easy substitution of new algorithms into an existing pipeline. Once written, a Titan component is immediately usable in programs written in C++, Python, TCL, and Java, allowing researchers to quickly deploy algorithms within the language that best suits their needs. Titan components can also be packaged into runtime "plugins" for use with OverView – Titan's general-purpose informatics application.

Georeferenced Semantic Graphs

Titan includes support for interactive displays of geo-referenced data on a zoomable, rotating globe. Background images and political boundaries can be displayed with dynamically-varying levels of detail, and georeferenced semantic graphs can be displayed in a variety of styles including support for nonoverlapping parallel edges. The combination of semantic graphs and geographic context can provide analysts with new views and new insights into existing data. As an example, network intrusion analysts are now using Titan to visualize geographic distributions of network traffic.

Parallel Graph Analysis

Special adapters allow the Titan graph data structures to be used with the MTGL graph library developed at SNL, which provides a variety of large-scale multi-threaded graph algorithms that can be run on both general- and special-purpose hardware. Similar adapters allow Titan to be used with the many high-quality serial graph algorithms provided by the Boost Graph Library, and work is ongoing with Indiana University to support distributed graph data structures including integration of the Parallel Boost Graph Library. When complete, the distributed graph data structures will allow Titan to be used for the deployment of parallel graph algorithms across a wide variety of HPC platforms.

Multi-dimensional Analysis

Titan provides a flexible set of sparse and dense n-way array data structures, allowing multi-dimensional data to be analyzed via algebraic methods. The PARAFAC tensor decomposition algorithm from the Tensor Toolkit is the first such algorithm integrated into Titan - future work will include integrating a distributed-memory parallel implementation of PARAFAC, additional decompositions such as TUCKER and DEDICOM, and integration with other libraries including Trilinos. This capability is currently being used to analyze terms and concepts appearing in network packet content.

Unstructured Text Analysis

Titan components for unstructured text analysis can perform binary and conceptual searches of an arbitrary corpus of text documents, along with global analysis of a corpus using conceptual similarity metrics. A new application - LSAView - provides visual comparisons of the impact of parameter changes on an analysis. Future work will include parallel distributed-memory analysis of extremely large corpora and relevance feedback techniques.

Statistical Analysis

Several OVIS project algorithms including descriptive, order, correlative, and contingency statistics have been integrated into Titan, where they can be applied to semantic graphs in novel ways. As

an example, statistical analysis is being used by network intrusion analysts to rapidly identify hosts whose network traffic patterns make them outliers.

Titan Resources

Website: http://vizrd.srn.sandia.gov/vizwiki/

Titan Mailing list: titan-users@mailgate.sandia.gov

Support e-mail: titan-help@sandia.gov

Tim Shead (1424) - tshead@sandia.gov

Brian Wylie (1424) - bnwylie@sandia.gov

John Greenfield (9326) - jagreen@sandia.gov

A.10 Physics Insights for Modeling and Data Analysis

Author: Jackson R. Mayo (8963)

Theoretical physics has systematically organized vast amounts of empirical data into a small number of simple, fundamental laws. Although many real-world problems are too complex for the laws of physics to be applied directly, physics offers guidance for model formulation and data analysis. Statistical mechanics and quantum field theory indicate that the effective dynamics of complex systems may share universal features across different application domains. Physics also teaches the importance of symmetry, dimensional analysis, and a flexible view of probability. Examples from turbulent combustion and from software performance optimization illustrate the application of such ideas to modeling.

For vital questions about the interpretation of observed data, physics is relevant in two different ways. First, the phenomena of interest are ultimately subject to basic physical laws. The application of these laws can involve *ab initio* models that predict a system's detailed behavior directly—as well as common-sense physics, such as the constraint that an effect cannot precede its cause, or the assumption that two data streams are statistically independent if they come from separated sources with no significant means of interaction. The fundamental laws of physics by themselves, however, are in practice insufficient for modeling many important systems, due to the enormous number of degrees of freedom.

The second role of physics, then, is to offer techniques, precedents, and inspiration for the construction of effective, non-fundamental models. This higher-level modeling is a major activity of physicists, such as in condensed matter physics, which studies substances with a complex, strongly interacting structure that usually cannot be analyzed by brute force. A major achievement of 20th-century physics was the recognition that higher-level reduced models often have a simple mathematical form resembling that of the fundamental laws, and sometimes can be derived systematically from those laws. This insight, in part, led to the successful application of physics methods to non-traditional fields such as traffic analysis, finance, and linguistics, which appear frequently in today's physics journals. In these fields, the principles and strategies of physics are arguably more valuable than the specific content of existing physics.

Almost all practical data analysis employs some kind of reduced model; even models accepted as complete descriptions of a system are usually well-founded abstractions of deeper processes. A useful framework in physics for simplifying the description of intractably large systems is renormalization. Many of the original degrees of freedom are omitted by sampling or averaging; those associated with the large-scale phenomena of primary interest are retained. The method is successful when the effect of the omitted degrees of freedom is approximately equivalent to that of substituting new values for the model parameters; these parameters are thus "renormalized" in the reduced description. Renormalization is useful when the underlying interactions exhibit locality, so that a block of nearby degrees of freedom behaves approximately as a unit—allowing a summary description of its internal state and higher-level interactions. Renormalization methods, originally motivated by multiscale fluctuations in statistical mechanics, have been applied extensively to similar phenomena, as in high-energy particle physics and fluid turbulence. A successful prediction

of these methods is that totally different underlying structures can lead to the same "universality class" of large-scale behavior; this, in turn, motivates the use of physics-inspired models even for non-traditional systems.

The most central concept in theoretical physics, one that can guide both model construction and data analysis, is symmetry (or invariance). The property that a system's behavior is unaffected by particular transformations, geometric or otherwise, can be inferred from its basic structure and imposes important constraints on the meaningful interpretation of data, limiting the range of models to be considered. Examples are the homogeneity and isotropy of space, which entail invariance to the choice of coordinate axes, and dimensional analysis, which entails invariance to the choice of units for physical quantities. Due to variations in measurement conventions, preprocessing, etc., data with the same meaning may appear as different numbers, and ideally an analysis will give results whose meaning is not affected by these variations. For example, when possible, general-purpose analysis methods should uphold a symmetry under independent rescaling of each variable; methods that use an *a priori* Euclidean metric in the data space may be imposing a meaningless relation between incommensurable quantities.

The confrontation between theory and experiment that drives physics has required the careful quantification of uncertainty from many sources, using the framework of probability. Understanding uncertainty is especially crucial for real-world applications where high-stakes decisions are guided by data analysis. Probabilistic elements in physics include theoretical uncertainty, in which a relevant mathematical fact (such as a high-order term in a series) has not yet been ascertained; epistemic uncertainty, in which unmeasured (but in principle measurable) influences affect an empirical result; pragmatic uncertainty, in which a theoretically deterministic process is so complex (chaotic) as to be random for all practical purposes; and intrinsic uncertainty, in which prediction of the outcome is deemed absolutely impossible (as in the standard interpretation of quantum mechanics). A key insight of physics is that, because different types of uncertainty are subject to the same mathematics of probability, they often can be treated interchangeably and can be modeled by one another. For example, statistical physics routinely treats pragmatic uncertainty as if it were intrinsic uncertainty. This flexibility enhances the ability of models to include all relevant uncertainty sources.

Work at Sandia's Combustion Research Facility on the foundations of turbulent combustion modeling illustrates some the concepts discussed. A thin flame sheet in a turbulent premixed fluid propagates locally at a speed determined by chemistry, while being wrinkled globally by flow vortices. The wrinkling increases the flame surface area and thus enhances the overall burning rate (speedup). Empirical measurements of the speedup as a function of turbulence intensity show wide scatter. Dimensional analysis indicates that the speedup can also depend on certain additional parameters of the flow, such as Reynolds number. In the Sandia work, a physics approach was taken by examining a regime of the problem with simpler behavior (weak turbulence) to provide a starting point for more general modeling. As in other work, turbulence (which is normally considered a deterministic but chaotic process) was modeled by an intrinsically random flow field, allowing use of various methods of statistical physics and quantum field theory. In the weak-turbulence limit, a mapping was identified that related the speedup to a property of a physics "toy model" called the directed polymer. Results from that model were translated and generalized to successfully predict the quantitative dependence of the speedup on Reynolds number, with implications even for the

strong-turbulence regime. Interestingly, the physics used was quite distinct from the lower-level physical laws that are normally considered to govern turbulent combustion (fluid mechanics and chemical kinetics).

Collaborative work between the OVIS project and the Scalable Computing R&D Department on software performance optimization gives another example of physics-inspired modeling. The available data consist of numerous low-level performance metrics, such as cache-miss counters, that are accumulated during an application run. It is difficult for a programmer to identify the performance bottlenecks and fruitful optimization strategies given this raw profile. In the approach under development, basic reference cases called kernels are run to establish a database of metric values characterizing specific performance issues. Data from application runs are transformed into a weighted average of the kernels, with the highest weights indicating the key performance patterns to be addressed for possible improvement. The method has shown promise by accurately diagnosing a known performance issue in an old version of the Babel software. Regression modeling is used to accommodate both epistemic uncertainty (the imperfect correspondence between kernels and real applications) and pragmatic uncertainty (fluctuations in event counts due to machine state). The linear structure of the model ensures that the diagnosis is invariant to arbitrary affine transformations of the vector of metrics (e.g., a redefined counter that always triggers twice as often, or a different grouping of events that causes an additional A count for every B count), as long as the kernels and the target application are profiled consistently. With this invariance comes the ability to extract information that is hidden in arbitrary linear combinations of the metrics. The use of simple and general analysis methods on a wide range of measured variables, whose significance may be unclear, is an efficient way to identify latent predictive relations.

DISTRIBUTION:

1 MS 0899 Technical Library, 8944 (electronic)

Sandia National Laboratories