Workshop Report: The Fourth Workshop on Mining Scientific Datasets, August 2001

Chandrika Kamath
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
Livermore, Ca 94551
kamath2@llnl.gov

The Fourth Workshop on Mining Scientific Datasets was held on August 26, 2001 in San Francisco, in conjunction with the Seventh SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001). The goal of the workshop was to bring together data miners analyzing science data and scientists from diverse fields to share their experiences, learn how techniques developed in one field could be applied to another, and understand some of the newer techniques being developed in the KDD community. This workshop, the fourth in the series, was also organized to better represent the significant amount of work that is taking place in the context of science and engineering applications; work that is sometimes overlooked in the mainstream KDD conferences. Earlier workshops in the series focused on the challenges in mining scientific data sets, identified the main disciplines that could help address these challenges, and discussed how the research being done in these diverse technical areas could be effectively harnessed to solve the problems of scientific data analysis.

The 10 talks at the fourth workshop included both the application of data mining to traditional scientific data sets such as those from remote sensing and astronomy, as well as the application of these techniques to new types of data such as mesh data from simulations. For example, Rie Honda from Kochi University in Japan, described the work she and her colleagues are doing in mining topographic features in large-scale planetary images. They focused on two difficulties in feature detection – the heterogeneity of image quality due to differences in illumination and surface conditions, and the wide range of target feature sizes. Using crater detection as a test-bed application, their approach used a combinatorial Hough transform along with genetic algorithms and preprocessing to reduce noise in the images. To handle the variation in image quality, they used Kohonen's self organizing maps to first create a rough grouping of images with similar quality. Then, using a representative image from each group, they fine tuned the parameters for the crater detection algorithms for the images in that group. While their approach performed better than methods using the FFT power spectrum, they believe that the accuracy of crater detection can be improved further by using other methods. A different view of the application of data mining in a traditional domain was given by Rahul Ramachandran from University of Alabama in Huntsville, who described the system architecture issues in the ADaM (Algorithm Development and Mining) system. He described the approach they have taken to address design problems in the mining of earth science data such as the variability of data sets, the diversity of operations for extracting information, and the capability to write complex mining plans. This talk provided useful insight into a topic rarely discussed in the KDD community, namely, how does one design a data mining system that can not only handle the variability in data sets and algorithms that are currently being used, but is also flexible enough to process new data sets and incorporate new algorithms without too much effort.

Two other talks in the workshop also focused on earth science data, analyzing the data from climate simulations and observations. Michael Steinbach, from the University of Minnesota, discussed the work he is doing with colleagues from NASA Ames Research Center, California State University, Monterey Bay, and University of Minnesota. By mining data from global observing satellites, terrestrial observations, and ecosystems models, they hope to find "interesting" patterns that could be used by climate scientists to better understand and predict changes in the global carbon cycle and climate system. Michael described the techniques they are using to handle spatio-temporal issues such as the removal of seasonality in order to detect non-seasonal patterns. Their early work has shown that by using a simple clustering algorithm such as K-means, and taking linear correlation as a measure of similarity between time series, it is possible to find ecosystem patterns that are already well known. While several research directions are being explored to improve their analysis approach, they hope that their close collaboration with earth scientists will help them to further investigate previously unknown patterns detected in the data. A different application of clustering in earth sciences was discussed by Scott Gaffney, a graduate student at UC Irvine. Working with simulation data, their goal is to cluster extra-tropical cyclone trajectories. Since these trajectories typically have different lengths, and include information about locality in space and time, their approach uses mixtures of regression models instead of feature-vector methods. First, the cyclones are tracked using the minimum of the mean sea-level pressure simulated every 6 hours. Next, these tracks are clustered using linear regression mixture components. Preliminary results indicate three clusters of tracks, with an additional background cluster that picks up tracks that appear not to belong to any other cluster.

A different application of data mining related to computer simulations was presented by Ramdev Kanapady, a graduate student at the University of Minnesota. One of the first steps in a computer simulation of a physical problem is the generation of a mesh to discretize the continuous domain. Typically, the process starts with a coarse mesh that is iteratively refined until results with acceptable accuracy are obtained. This process is both cumbersome and expensive. The approach proposed in this work was to extend previous work that had used neural networks to predict mesh density in computational electro-magnetics, to the problem of predicting mesh density in structural mechanics applications. Using a simple test problem of a square plate, simply supported at the edges, with a concentrated load applied at a known point, Kanapady and his colleagues showed that it was possible to predict close to ideal mesh density over the plate. In this simple proof-of-concept application, the analytical solution was known and therefore could be used as a training set. However, one of the challenges in extending this work to a general problem is to find a strategy for generating a training set. This lack of labeled examples is a common problem in scientific data mining; one approach to addressing it in this particular example would be to use meshes generated by experts.

Prediction of damage was another topic of discussion in the workshop. John Brence described the work he has been doing at the University of Virginia, Charlottesville, in predicting corrosion in commercial and military aircraft so that corroded parts can be replaced at an appropriate time. Using non-destructive test data, they tried different methods, such as multiple linear regression, regression trees, logistic regression, and polynomial networks to accurately predict corrosion. They found that the more complex models did better on training data, but poorly on test data. Early results indicated that the least absolute deviation regression tree model worked the best. More importantly, use of data mining in this application could complement the identification of corrosion by trained operators to increase the reliability of the process. A different aspect of damage detection was presented by Sukhpreet Sandhu, a graduate student from the University of Minnesota. The focus of his work was to use data mining to identify a damaged element in the simulation of a structure. Using simple structures, damage in an element was simulated by reducing its Young's modulus. Using features such as the absolute static displacements of the nodes, and decision trees and neural networks to build the models, the task was to identify the element that had failed. This work considered both simple structures with constant and variable loading, as well as more complex structures such as a transmission tower. Preliminary results showed that in order to build an accurate model, features that had a close relationship with the target function must be extracted from the simulation data. It is expected that this requirement will become more important as we try to accurately predict the damage in complex structures.

Other talks in the workshop included the use of genetic programming to predict of the performance of engine oil using test data by Tina Yu from Chevron, and the discovery of interesting rules in critical care databases, by Jafar Adibi, a graduate student at University of Southern California. A novel application of data mining was discussed by Dennis DeCoste from the Jet Propulsion Laboratory, who described the work he is doing with colleagues at CalTech on analyzing the data from an "electronic nose". Here, an array of polymer films embedded with conductive or resistive material is exposed to a vapor, leading to a change in the electrical resistance of the films. This data, after some pre-processing, can be analyzed to classify the vapor into its chemical group. DeCoste presented the performance of support vector machines and kernel Fisher discriminants in accurately identifying the chemical category of a vapor. Their early results show that kernel methods offer some promise for this real-world task, though more work is needed to accurately and fairly identify the best kernel.

The 10 talks in the workshop described several practical applications of data mining in scientific and engineering domains. They also illustrated some of the problems that need to be addressed such as the identification of more representative features for a problem, more robust ways of extracting features, the ability to handle variations in data, and better ways of generating training sets. In addition to the application of data mining to traditional domains such as astronomy and remote sensing, several of the talks also described early work in newer application areas such as analysis of non-destructive test data and simulation data. Several of the talks were presented by graduate students, indicating an increasing interest in an emerging field.

The co-organizers of the fourth workshop were Michael Burl (Jet Propulsion Laboratory), Chandrika Kamath (Lawrence Livermore Laboratory), Vipin Kumar (University of Minnesota), and Raju Namburu (Army Research Laboratory). The next workshop in this series will be held in conjunction with the second SIAM International Conference on Data Mining in April 2002. Additional details on this, and earlier workshops, are available at

http://www.ahpcrc.umn.edu/conferences/

The talks presented at the first two workshops in this series have been collected in a book "Data Mining for Scientific and Engineering Applications", edited by Robert L. Grossman, Chandrika Kamath, Philip Kegelmeyer, Vipin Kumar, and Raju R. Namburu, and published by Kluwer Academic Publishers in September 2001.

Acknowledgement

The work of Chandrika Kamath was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W7405-Eng-48.

About the author:

Chandrika Kamath is the project lead for Sapphire, a project in large scale data mining and pattern recognition at the Center for Applied Scientific Computing at the Lawrence Livermore National Laboratory (http://www.llnl.gov/casc/sapphire). She received the B.Tech degree in Electrical Engineering from the Indian Institute of Technology, Bombay, in 1981, and the Master's and Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign in 1984 and 1986, respectively. Her current interests are in all aspects of data mining, including image processing, feature extraction, dimension reduction, and classification and clustering algorithms. She is also interested in the practical application of these techniques to scientific data sets that result from observations, experiments, and simulations.