Report on DIMACS Tutorial on Data Mining and Epidemiology

Dates: March 23 - 24, 2006

Location: DIMACS Center, CoRE Building, Rutgers University

Organizers:

James Abello, DIMACS and Ask.com; Graham Cormode, Bell Labs

(DIMACS is a consortium of Rutgers and Princeton Universities, AT&T Labs, Bell Labs, NEC Laboratories America, and Telcordia Technologies, with partners Avaya Labs, IBM Research, and Microsoft Research)

1. Introduction

Data Mining is now a staple part of Computer Science, and has been applied in a wide variety of different areas. It covers a diverse set of topics from algorithms, statistics and discrete mathematics, with the general goal of identifying patterns in data in order to draw inferences and make predictions. This tutorial brought together experts from Data Mining to introduce the key ideas and techniques from:

- Probability, Decision Trees and Bayesian Statistics
- Machine Learning, Classifiers and Boosting
- Data Stream Analysis and Clustering
- Graph Mining
- Applications to Biology and Epidemiology

The goal was to allow people with little or no knowledge of data mining to understand the basic techniques, and get a flavor of the general methodology and style of results. This tutorial was aimed to be of interest to researchers wishing to work in data mining, and also to researchers from outside computer science who wish to understand these methods in order to apply them. The tutorial included short talks on applications to problems in epidemiology and biology in order to put the general techniques described into perspective.

In detail, the talks and tutorials presented were as follows:

2. Tutorials

Probability for Data Miners

Brigham Anderson, Carnegie Mellon University

Probability lies at the heart of data mining. This tutorial introduced these basic probabilistic foundations, with an emphasis on Bayesian concepts. The axioms of probability were reviewed, and then these were used to create probabilistic models, which underlie many machine learning algorithms. The simplest models were discussed: the full joint density and the naïve density, and it was shown how they can be used to perform the common tasks of machine learning: inference, classification, anomaly detection, and clustering. Along the way, simple Bayesian networks were used to characterize the models and assumptions. Concepts discussed included Sample spaces, Bayes rule, conditional independence, Bayes classifiers, naïve Bayes classifiers, Bayes nets, hidden variables, generative models, and mixture models.

Machine Learning Algorithms for Classification

Robert Schapire, Princeton University

Machine learning studies the design of computer algorithms that automatically make predictions about the unknown based on past observations. Often, the goal is to learn to categorize objects into one of a relatively small set of classes. This tutorial introduced some of the main state-of-the-art machine learning techniques for solving such classification problems, namely, decision trees, boosting, support-vector machines and neural networks. The tutorial also discussed some of the key issues in classifier design, including avoidance of overfitting.

Cluster and Data Stream Analysis

Graham Cormode, Bell Laboratories

Clustering is an important tool in machine learning and data mining. It allows features and correlations in the data to be identified and requires few parameters and little detailed information about the data. The results can be used to generate hypotheses, aid in visualization, or reduce the data to a few prototypical points. This 'unsupervised learning' technique has many variants and many perspectives. This tutorial gave an algorithmic view, describing some of the most popular clustering algorithms and identifying their pros and cons, including hierarchical clustering, k-means, expectation maximization (EM) and k-center approximation algorithms. When the input data is too large to conveniently hold in memory, or is being constantly updated, it is necessary to view the data as a massive stream. In recent years the "data stream" model has become a popular way to handle massive data sources. The tutorial outlined some of the key properties of data streams, and illustrated them with some of the recent work in clustering on data streams.

Graph Mining

James Abello, DIMACS / Rutgers University and Ask.com

A variety of massive data sets exhibit an underlying structure that can be modeled as dynamic weighted multi-digraphs. Their sizes range from tens of gigabytes to petabytes. These include the World Wide Web, Internet Traffic and Telephone Call Detail. These

data sets sheer volume brings with it a series of computational and visualization challenges due mainly to the I/O and Screen Bottlenecks. The tutorial presented external memory algorithms for connectivity and minimum spanning trees together with heuristics for quasi-clique finding. It was described how hierarchy trees help us to cope in a unified manner with both, the I/O and screen bottlenecks. This line of research has suggested the need to look for "novel" graph representations in order to provide a user or a data-mining engine with informed navigation paths. Results were shown with graphs having on the order of 200 million vertices and several billion edges and mentioned some mathematical problems that have surfaced along the way. The overall goal is to extract useful information that can be brought into a user's palm top and to export these techniques to other mining domains.

3. Other Contributed Talks

The Containment Problem

Michael Capalbo

Consider the following problem, known as the Containment Problem. Let G be a network (e.g., social, computer network, etc), and let S_0 be any subset of the nodes of G, such that every node in S_0 is infected with a virus that spreads from each infected node to all of its non-vaccinated neighbors in one time step. The allowed response is to vaccinate a limited number a_l of nodes during each time step l=1,2, ... t, for some integer t. The Containment Problem asks us to find which a_l nodes to vaccinate at each time step l to minimize the total number of nodes that eventually become infected, or in other words, devise an optimum vaccination strategy given our resources. The Containment Problem has applications in, for example, Computational Epidemiology. Unfortunately, for general instances of the Containment Problem, there is no known tractable algorithm for returning an optimum vaccination strategy. The focus of this talk was to present a tractable algorithm that returns a vaccination strategy that is only slightly inferior to the optimum vaccination strategy (under certain assumptions on G).

Viruses and Computer Scientists

Courtney D. Corley, University of North Texas

Recent world events such as the threat of bioterrorism, the outbreak of SARS and the spread of the H5N1 Avian Influenza have motivated computer scientists to begin looking at issues in the public health and epidemiology domains. This talk summarized current, exciting research in modeling and simulating the spread of viruses. It also introduced basic knowledge about the structure and function of viruses, as well as discussing some of the epistemology surrounding whether viruses are "Alive" or "Dead".

Selected Problems in Epidemiology

Nina H. Fefferman, DIMACS and Tufts University

Recent advances in technology have led to more accurate and more complete methods of bio-surveillance. As a result, finding useful information within the sheer amount of data

gathered has become difficult. Modern problems in public health and epidemiology require better techniques in data mining in order to make sense from the chaos. This talk discussed a few examples from U.S. public health and further investigated some of the trade-offs in monitoring necessitated by a lack of appropriate data mining methods.

Using cluster analysis to determine the influence of epidemiological features on medical status of lung cancer patients

Dmitriy Fradkin, Ask.com

This work analyzed lung cancer data, obtained from SEER, for 217,558 patients diagnosed in 1988-2000. Each patient is characterized by 23 epidemiological (essentially demographic) and 22 medical features. The main idea of this analysis consists in clustering the data in the space of epidemiological features only, and analyzing influence of the epidemiological classification on medical status of patients. The influence is estimated by using the T-test to determine differences in the distributions of medical features between clusters. The epidemiological part of data was partitioned into 20 clusters. Out of 190 cluster pairs, there are 2 pairs with only 1 distinguishing medical feature and 4 pairs with 2 distinguishing features. All other pairs differ in at least 3 medical features. Some medical features are not different in any pair of clusters, and some take distinct values in many clusters. Such analysis indicates which medical aspects are most affected by epidemiological status. On the other hand, it aids in finding epidemiological subpopulation (clusters) that are very different from others in their medical characterization.

Mathematical Formulation of the Foot-and-mouth Disease Epidemic Component of the Decision Support System Developed at LLNL

Tanya Kostova, LLNL

Development of a decision support system to evaluate the spread and economic impact of a possible introduction of foot and mouth disease in the US is under way at Lawrence Livermore National Labs (LLNL). The model is an agent-based stochastic simulation. The agents are the individual animal facilities (farms, feedlots, sales-yards, etc.) and they can be thought of as the nodes of a network. The nodes can be infected via several methods and pass through the commonly accepted stages of infection (latent, subclinical, clinical, immune, etc.). This talk presented the basic mathematical formulation of the model, the type of input the model uses and the type of simulation output we will be obtaining. The code is under construction.

Re-interpreting DNA Microarray Data

Sungchul Ji, Rutgers University

In recent years, data obtained from microarrays has been analyzed by a variety of techniques, including clustering and visualization. This talk sought to explain the methodology of such analysis, and identify some of the shortcomings of this understanding, based on misconceptions that have been propagated in the literature many times over.

4. Conclusions

The tutorial attracted a large amount of interest from potential attendees, with over 80 registering an interest. Of these, over sixty participants attended the tutorial and participated in discussions. The feedback on the event was very positive, with extra praise directed towards the tutorial component of the meeting. Slides from all talks and tutorials are available freely from the tutorial website, at:
http://dimacs.rutgers.edu/Workshops/DataMiningTutorial/slides/slides.html

Acknowledgements

The organizers and the DIMACS Center acknowledge the support of the National Science Foundation under grant number CCR 03-14161 to Rutgers University.

James Abello
Graham Cormode