DIMACS Working Group on Data Mining and Epidemiology

Dates: First meeting May 22 - 23, 2003
DIMACS Center, CoRE Building, Rutgers University

Organizers:
Ilya Muchnik, Rutgers University, muchnik@dimacs.rutgers.edu

S. Muthukrishnan, AT&T Labs - Research and Rutgers University, muthu@cs.rutgers.edu

David Ozonoff, Boston University, dozonoff@bu.edu

Presented under the auspices of the Special Focus on Computational and Mathematical Epidemiology. New computational methods are needed to deal with large, complex data sets arising in epidemiology. In this working group, epidemiologists will join with computer scientists and statisticians to explore new uses of data mining methods in epidemiology. Notifiable infectious diseases provide a huge testbed of data for surveillance, for planning and evaluation of intervention programs, and for hypothesis generation in etiologic studies [Levin, Grenfell, Hastings and Perelson (1997)]. Data sets connecting environmental factors and disease can be used in much the same way [Paulu and Ozonoff (1998), Vieira, Webster, Aschengrau and Ozonoff (2001)]. Some examples of large data sets arising in epidemiological studies involve fat-free body mass [Chagnon, Y.C., Borecki, Prusse, Roy, Lacaille, Chagnon, M., Ho-Kim, Rice, Province, Rao and Bouchard (2000),], linkage scans [Province and Single (2000)], the health of Gulf War veterans [Proctor, Heeren, White, Wolfe, Borgos, Davis, Pepper, Clapp, Sutker, Vasterling and Ozonoff (1998)] and alcoholism [Saccone, Kwon, Corbett, Goate, Rochberg, Edenberg, Foroud, Li, T., Begleiter, Reich and Rice (2000),]. A long list of links to large health-related data sets can be found at the website http://www.ehdp.com/vitalnet/datasets.htm. Such data, while often massive in quantity, is uneven in quality and completeness and heterogeneous in nature. This group will build on relevant data mining studies in the epidemiological literature (see e.g., [Brossette, Jones, Sprague, Hardin and Moser (1999), Brossette, Sprague, Hardin, Waites, Jones and Moser (1998), DuMouchel (1999), Forgionne, Gangopadhyay and Adya (2000), Holmes, Durbin and Winston (2000), Openshaw, Turton and MacGill (1999), Pendharkar, Rodger, Yaverbaum, Herman and Benner (1999), Richards, Rayward-Smith, Sonksen, Carey and Weng (2001)]). We will emphasize the development of new algorithmic methods for data mining in epidemiology involving visualization, clustering, and aggregation. Automatic environmental monitoring and risk evaluation for cancer provides a sample motivation. In the US, there is a rich history of cancer mapping, highlighted by the release of the first US Cancer mortality atlas in 1975, the recent development of the US Cause of Death Atlas, and the National Cancer Institute's data set of about 10 million US cancer cases. Additional data comes from questionnaires based on individual patient and resident information; public registries with cancer incidences aggregated by county; population-based cancer registries aggregated by city and town; birth and death registries; environmental data such as sample databases of water conditions and air quality records; census data such as geographic databases with accurate locations of population; and

remotely-sensed data providing information on land use patterns or air pollution distribution. All of these databases have different temporal and spatial assumptions (for example, different frequencies of collection, different spatial resolution (by state, by county, by zip-code, by square kilometer), etc. Cluster analysis offers the promise of pattern extraction from such complex data. Our approach to clustering will start by emphasizing data cleaning tools (see, e.g., [Galhardas, Florescu, Shasha and Simon, E. (2000), Galhardas, Florescu, Shasha and Simon, E. (2000), Lambert, Pinheiro and Sun (1999)] and the website http://www.research.att.com/~tamr/dataquality.html) since so much epidemiological data has problems arising from manual entry, lack of uniform standards for content and formats, data duplication, and measurement errors. We shall investigate such solutions as duplicate removal, merge purge, and automated detection. Application of traditional clustering algorithms is hindered by the extreme heterogeneity of the data and we shall discuss new approaches to deal with such heterogeneities. Promising algorithmic methodologies for clustering heterogeneous data are in the papers [Kempner, Mirkin and Muchnik (1997), Kuznetsov and Muchnik (1982), Muchnik and Shvartser (1990)]. We shall also build on traditional statistical approaches to heterogeneous epidemiological data (see, e.g., [Cox and Piegorsch (1996), Dominici, Parmigiani, Wolpert and Hasselblad (2000), Patil (1991)]). Huge data sets are sometimes best understood by visualizing them. Sheer data sizes require new visualization regimes, which require suitable external memory data structures to reorganize tabular data in secondary storage so that access, usage, and analysis are facilitated (see, e.g., [Abello (1999)]). The working group will be faced with the challenge that developing visualization algorithms becomes harder when data arises from various sources and each source contains only partial information, as is the case, for example, with the cancer monitoring data. In cancer monitoring, we start with data about individuals, regions, etc., and seek to produce aggregative indices, single or multivariate numerical values that measure the risk that an entity in a given group will come down with a certain form of the disease. The groups consist of sets of individuals in different geographic regions, under different environmental conditions (known contaminated drinking water, nearness to environmental risks, etc.), or with different individual risk factors (age, smoking, etc.). While worrying about the "meaningless" statements one can make using such aggregative indices ([Roberts (1994), Roberts (1999)]), we will seek to develop methods for obtaining such indices that are both efficient to compute and useful as predictive tools. We will build on recent approaches to this problem in the bioinformatics context [Hagerty, Muchnik, Kulikowski and Kim (1999)] and in the information technology context [Pennock, Maynard-Reid, Giles and Horvitz (2000), Schapire, Freund, Bartlett and Lee (1998)]. These methods use learning algorithms to develop aggregations using individual "classifiers" based on different sources of data and a "compromiser" to merge the results. This is one of several working groups that will benefit from the involvement of researchers from DIMACS' industrial partners who will help us modify for epidemiological purposes the methods they have developed for telecommunications/computer network applications - a unique aspect of this project.

References:

Abello, J.M., and Vitter, J.S. (1999), *External Memory Algorithm*s, DIMACS Series, vol. 50, American Mathematical Society, Providence, RI.

Brossette, S.E., Jones, W.T., Sprague, A.P., Hardin, J.M., and Moser, S.A. (1999), "DMSS: A knowledge discovery system for epidemiologic surveillance," *The International Journal of Knowledge Discovery and Data Minin*g.

Brossette, S.E., Sprague, A.P., Hardin, J.M., Waites, K.B., Jones, W.T., and Moser, S.A. (1998), "Association rules and data mining in hospital infection control and public health surveillance," *Journal of the American Medical Informatics Associatio*n, 5, 373-381.

Chagnon, Y.C., Borecki, I.B., Prusse, L., Roy, S., Lacaille, M., Chagnon, M., Ho-Kim, M.A., Rice, T., Province, M.A., Rao, D.C., and Bouchard, C. (2000), "Genome-wide search for genes related to the fat-free body mass in the Quebec family study," *Metabolis*m, **4**9, 203-207.

Cox, L.H. and Piegorsch, W.W. (1996), "Combining environmental information I: Environmental monitoring, measurement and assessment" and "Combining environmental information II: Environmental epidemiology and toxicology," *Environmetric*s, 7, 299-308 and *Environmetric*s, 7, 309-324.

Dominici, F., Parmigiani, G., Wolpert, R. and Hasselblad, V. (2000), "Meta-analysis of migraine headache treatments: Combining information from heterogeneous designs," *Journal of the American Statistical Associatio*n, **9**4, 16-28.

DuMouchel, W. (1999), "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system, with discussion," *The American Statistician*, **5**3, 177-202.

Forgionne, G.A., Gangopadhyay, A., Adya, M. (2000), "Cancer surveillance using data warehousing, data mining, and decision support systems," *Top. Health Inf. Manage.*, **2**1, 21-34.

Galhardas, H., Florescu, D., Shasha, D., and Simon, E. (2000), "An extensible framework for data cleaning," *Proceedings of the 16th Intern. Conf. on Data Engineering* (ICDE), 312.

Galhardas, H., Florescu, D., Shasha, D., and Simon, E. (2000), "AJAX: An extensible data cleaning tool," *SIGMOID Conferenc*e, 590.

Hagerty, C.G., Muchnik, I., Kulikowski, C., and Kim, S-H. (1999), "Two indices can approximate four hundred and two amino acid properties," *Proceedings of the IEEE International Symposium on Intelligent Contro*l, Cambridge, MA, 365-369.

Holmes, J.H., Durbin, D.R., and Winston, F.K. (2000), "Discovery of predictive models in an injury surveillance database: An application of data mining in clinical research," *Proc AMIA Sym*p, 359-363.

Kempner, Y., Mirkin, B., and Muchnik, I. (1997), "Monotone linkage clustering and quasi-concave set functions," Applied Mathematics Letters, 4, 19-24, http://www.datalaundering.com/download/kmm.pdf.

Kuznetsov, E., and Muchnik, I. (1982), "Analysis of the distribution of functions in an organization,"Automation and Remote Control, 43, 1325-1331, http://www.datalaundering.com/download/organiza.pdf .

Lambert, D., Pinheiro, J., and Sun, D. (1999), "Reducing transaction databases, without lagging behind the data or losing information," Technical Report, Statistics and Data Mining Research Department, Bell Labs, Lucent Technologies, Murray Hill, NJ.

Levin, S.A., Grenfell, B., Hastings, A., and Perelson, A.S. (1997), "Mathematical and computational challenges in population biology and ecosystems science," *Scienc*e, **1**7, 334-343.

Muchnik, I., and Shvartser, L. (1990), "Maximization of generalized characteristics of functions of monotone systems," Automation and Remote Control, 51, 1562-1572, http://www.datalaundering.com/download/maxgench.pdf.

Openshaw, S., Turton, I. and MacGill, J. (1999), "Using the geographic analysis machine to analyze limiting long-term illness census data," *Geographic and Environmental Modelin*g, 3, 83-99.

Patil, G.P. (1991), "Encountered data, statistical ecology, environmental statistics, and weighted distribution methods," *Environmetric*s, 2, 377-423.

Paulu, C., and Ozonoff D. (1998), "Exploring associations between residential history and breast cancer risk in a case-control study," *International Society for Environmental Epidemiology/International Society of Exposure Analysis, Annual Conferenc*e, Boston, MA, August 15-18, 1998.

Pendharkar, P.C., Rodger, J.A., Yaverbaum, G.J., Herman, H. and Benner, M. (1999), "Association, statistical, mathematical and neural approaches for mining breast cancer patients," *Expert Systems with Application*s, **1**7, 223-232.

Pennock, D.M., Maynard-Reid, P., Giles, C.L., and Horvitz, E., A. (2000), "A normative examination of ensemble learning algorithms," *Proc. 17th International Conference on Machine Learning (ICML-2000*), Morgan-Kauffmann, 735-742.

Proctor, S.P., Heeren, T., White, R.F., Wolfe, J., Borgos, M.S., Davis, J.D., Pepper, L., Clapp, R., Sutker, P.B., Vasterling J.J., and Ozonoff, D. (1998), "Health status of Persian Gulf War veterans: Self-reported symptoms, environmental exposures and the effect of stress," *International Journal of Epidemiology*, **2**7, 1000-1010.

Province, M.A., and Single, A. (2000), "Sequential, genome-wide test to identify simultaneously all promising areas in a linkage scan," *Genet. Epid*., **1**9, 301-322.

Richards, G., Rayward-Smith, V.J., Sonksen, P.H., Carey, S., and Weng, C. (2001), "Data mining for indicators of early mortality in a database of clinical records," *Artif Intell Me*d, **2**2, 215-231.

Roberts, F.S. (1994), "Limitations on conclusions using scales of measurement," in Barnett, A., Pollock, S.M., and Rothkopf, M.H. (eds.), *Operations Research and the Public Secto*r, Elsevier, Amsterdam, 621-671.

Roberts, F.S. (1999), "Meaningless statements," in Graham, R.L., Kratochvil, J., Nesetril, J., and Roberts, F.S. (eds.), *Contemporary Trends in Discrete Mathematic*s, DIMACS Series, **4**9, American Mathematical Society, Providence, RI, 257-274.

Saccone, N.L., Kwon, J.M., Corbett, J., Goate, A., Rochberg, N., Edenberg, H.J., Foroud, T., Li, T., Begleiter, H., Reich, T., and Rice, J.P. (2000), "A genome screen of maximum number of drinks as an alcoholism phenotype," *Am. J. Med. Genet*., **9**6, 632-637.

Schapire, R.E., Freund, Y., Bartlett, P., and Lee, W.S. (1998), "Boosting the margin: A new explanation for the effectiveness of voting methods," *The Annals of Statistic*s, **2**6, 1651-1686.

Vieira , V., Webster, T, Aschengrau, A, and Ozonoff, D. (2001), "A method for spatial analysis of risk in a population-based case-control study, *International Journal of Hygiene and Environmental Healt*h, submitted.