Interface 2004 Schedule

Wednesday, May 26, 2004

7:30 a.m. — Registration Begins (Harborside Coatroom)

8:00 a.m. to 5:30 p.m. — Interface Half-day Short Courses (Essex)

8:00 p.m. – 12:00 p.m. Short Course I: Random Forests, Leo Breiman and Adele Cutler

1:30 p.m. – 5:30 p.m. Short Course II: Gene Expression Analysis, Rafael A. Irizarry

6:00 p.m. to 8:00 p.m. — Interface Board of Governors Meeting (closed, Boardroom)

8:00 p.m. to 10:00 p.m. — Reception and Mixer (Harborside Ballroom D)

Thursday, May 27, 2004

7:30 a.m. — Registration Continues (Harborside Registration B)

8:30 a.m. – Opening Remarks (Harborside Ballroom A/B)

8:45 a.m. to 9:45 a.m. Keynote Address (Harborside Ballroom A/B): *Computational Biology and the Ribosome*, David E. Draper, Johns Hopkins University

9:45 a.m. to 10:30 a.m. – Coffee Break (Harborside Foyer)

10:30 a.m. to 12:15 a.m. – Tubes and Graphs (Harborside A — Invited Session)
Organizer: Wendy Martinez, Office of Naval Research
Session Chair: Wendy Martinez, Office of Naval Research

10:30 — *The Volume-of-Tube formula: Computational Methods and Statistical Applications.*, Catherine Loader, Case Western Reserve University, and Ramani S. Pilla, Case Western Reserve University

11:20 — *Multivariate Statistics for Trees and Networks*, Susan Holmes, Stanford University

---

10:30 a.m. to 12:15 a.m. – Internet Tomography (Harborside B — Invited Session)
Organizer: Steve Marron, UNC Chapel Hill
Session Chair: Amy Braverman, JPL

10:30 a.m. — *Wavelet and SiZer Analyses of Internet Traffic Data*, Cheolwoo Park, SAMSI

11:20 a.m. — *Tools for the Analysis and Visualization of Network Traffic*, George Michailidis, The University of Michigan

---

10:30 a.m. to 12:15 a.m. – Too Many Features, Too Few Samples – Challenges for Disease Profiling Using Biomedical Data (Essex — Invited Session)
Organizer: Ray Somorjai, National Research Council, Canada
Chair: Ray Somorjai, National Research Council, Canada

10:30 a.m. — *The Analysis of Biomedical Data - Caveats and Challenges*, R.L. Somorjai, Institute for Biodiagnostics, National Research Council Canada

11:05 a.m. — *Supervised Analysis When the Number of Features (p) Greatly Exceeds the Number of Cases (n)*, Richard Simon, National Cancer Institute

11:40 a.m. — *Many Features, Few Samples: From Cheminformatics to Bioinformatics*, Kristin P. Bennett, Rensselaer Poly. Tech., Curt Breneman, Rensselaer Poly. Tech, and Mark Embrechts, Rensselaer Poly. Tech.

10:30 a.m. to 12:15 a.m. – Visualization (Falkland — Contributed Session)

10:30 a.m. — *Extensions to Dynamically Conditioned Choropleth Maps*, Yuguang Zhang, FreddieMac, and Daniel B. Carr, George Mason University (refereed)

10:50 a.m. — *Visual Analytics for Dynamically Conditioned Choropleth Maps: QQ-plots, Scatterplot Smoothes and Two-Way Tables*, Chunling Zhang, George Mason University, Yaru Li, George Mason University, and Daniel Carr, George Mason University

11:10 a.m. — *Widget Management in the Multilayered 3-D Visual Analytic Software Called Glisten*, Yanling Liu, George Mason University, and Daniel Carr, George Mason University

11:30 a.m. — *Visualizing Patients Treated with Three-Dimensional Computed Tomography-Guided Brachytherapy* Faleh Alshameri, (George Mason University), and Jee Vang, (George Mason University)

11:50 a.m. — *Multidimensional Scaling and Classification*, Michael W. Trosset, College of William & Mary

---

10:30 a.m. to 12:15 a.m. – Microarrays I (Galena — Contributed Session)

10:30 a.m. — *Extending the Loop Design for Microarray Experiments*, Naomi S. Altman, Pennsylvania State University

10:55 a.m. — *Mixed Effects Model for Assessing RNA Degradation in Affymetrix GeneChip Experiments*, Kellie J. Archer, Ph.D., Department of Biostatistics, Center for the Study of Biological Complexity, Virginia Commonwealth University, Suresh E. Joel, Department of Biomedical Engineering, Virginia Commonwealth University, and Viswanathan Ramakrishnan, Department of Biostatistics, Virginia Commonwealth University

11:20 a.m. — *Performance of the False Discovery Rate for Small Sets of cDNA Microarrays*, Simon Rosenfeld, National Cancer Institute (refereed)

11:45 a.m. — *Mutual Information Based ICA Applied to Microarray Data*, Hoang, M. Thu, Universite Rene Descartes, Pham, Dinh Tuan, Centre National de la Recherche Scientifique, IMAG, and Thierry-Pascal Baum, Centre National de la Recherche Scientifique, IMAG

12:15 p.m. to 1:45 p.m. — Lunch Break

1:45 p.m. to 3:30 – Mixture Modeling: Modern Approaches and Applications (Harborside A — Invited Session)
Session Organizer: Richard Charnigo, University of Kentucky
Session Chair: Richard Charnigo, University of Kentucky

1:45 p.m. — *The Volume-of-Tube Formula: Applications to Perturbation and Mixture Models*, Ramani S. Pilla, Case Western Reserve University, and Catherine Loader, Case Western Reserve University

2:20 p.m. — *The Role of Latent Variables in Model Selection Accuracy*, Jeremy Nadolski, University of Kentucky

2:55 p.m. — *Empirical Likelihood Based Inferences in Semiparametric Finite Mixtures*, Jing Qin, Memorial Sloan-Kettering Cancer Center

---

1:45 p.m. to 3:30 – Analysis of Functional Neuroimaging Data (Harborside B — Invited Session)
Session Organizer: Bill Eddy, Carnegie Mellon University
Session Chair:

1:45 p.m. — *Brains on Film: Using Optical Imaging to Build Maps of Brain Activity*, Kary Myers, Carnegie Mellon University

2:20 p.m. — *Reducing Physiological Noise in fMRI Experiments*, Rebecca L. McNamee, University of Pittsburgh

2:55 p.m. — *Wavelet-Based Statistical Analysis of fMRI Data*, Ivo Dinov, UCLA Statistics/Neurology, Arthur Toga, UCLA Neurology, Michael Mega, Neural Net Research, and John Boscardin, UCLA Biostatistics

---

1:45 p.m. to 3:30 p.m. – Text Mining for Biomedical Applications (Essex — Invited Session)
Organizer: Lynette Hirschman, MITRE
Chair: Lynette Hirschman, MITRE

1:45 p.m. — *Natural Language Processing for Biosurveillance*, Dr. Wendy W. Chapman, University of Pittsburgh

2:10 p.m. — *Automated Terminological Networks For High Throughput Comparative Biology of Phenotypes*, Yves A. Lussier , Columbia University, and Xiaoyan Wang, Columbia University

2:35 p.m. — *The MiTAP System for Monitoring Reports of Disease Outbreak*, Laurie Damianos, The MITRE Corporation

3:00 p.m. — *Project Argus*, James Wilson, ISIS Center, Georgetown University

---

1:45 p.m. to 3:30 – Computational Geometry and Robust Statistics (Falkland — Invited Session)
Session Organizer: Diane Souvaine, Tufts University
Session Chair: Eynat Rafalin, Tufts University

1:45 p.m. — *On the Least Median Square Problem*, David M. Mount, University of Maryland, Jeff Erickson, University of Illinois, and Sariel Har-Peled, University of Illinois

2:20 p.m. — *Computational Geometry and Statistical Depth measures*, Eynat Rafalin, Tufts University, and Diane L. Souvaine, Tufts University

2:55 p.m. — *Imputation of Microarray Data*, Ming Ouyang, University of Medicine and Dentistry of New Jersey

---

1:45 p.m. to 3:30 – Microarrays II (Galena — Contributed)

1:45 p.m. — *Rank-Based Classification of Gene Expression Profiles*, Donald Geman, Daniel Naiman and Christian d'Avignon, Dept. of Applied Mathematics and Statistics, and Whitaker Biomedical Engineering Institute, Johns Hopkins University

2:10 p.m. — *A Comparison of Direct and Sequential False Discovery Rate Algorithms: Computational Experiments for Exploratory DNA Microarray Studies*, Danh V. Nguyen, University of California, Davis

2:35 p.m. — *Probabilities of Spurious Connections in Gene Networks: Application to Expression Time-series*, David R. Bickel, Medical College of Georgia, refereed

3:00 p.m. — *DNA Microbial and Viral Identification using Ultraspecific Probes "Blind" to Host and Background DNA*, Catherine Putonti, University of Houston, George E. Fox, University of Houston, Richard C. Willson, University of Houston, and Yuriy Fofanov, University of Houston

---

3:30 p.m. to 4:00 p.m. – Refreshment Break (Harborside Foyer)

4:00 p.m. to 5:45 – Functional Data Analysis for Computational Biology (Harborside A — Invited Session)
Session Organizers: Catherine Loader and Ramani Pilla, Case Western Reserve University
Session Chair: Catherine Loader, Case Western Reserve University

4:00 p.m. — *Self modeling with Flexible, Random Time Transformations*, Lyndia C. Brumback, University of Washington, and Mary J. Lindstrom, University of Wisconsin

4:35 p.m. — *Modeling Continuous Shape Change for Facial Animation*, Julian Faraway, University of Michigan

5:10 p.m. — *Detecting Changes in Brain Shape, Scale and Connectivity via the Geometry of Random Fields* , Keith Worsley, McGill University

---

4:00 p.m. to 5:45 – Future of Statistical Software (Harborside B — Invited Session)
Session Organizer: Jim Gentle, George Mason University
Session Chair:

4:00 p.m. — *Challenges for Future Statistical Software for Non-expert Users*, John Sall, SAS Institute

4:35 p.m. — *Yxilon – Designing the Next Generation, Vertically Integrable Statistical Computing Environment*, Uwe Ziegenhagen, HU Berlin, and Prof. Dr. W. Haerdle, HU Berlin

5:10 p.m. — *XML-Based Applications in Statistical Analysis*, Yuichi Mori, Okayama University of Science, Tomokazu Fujino, Fukuoka Women's University, Yoshiro Yamamoto, Tama University, and Tomoyuki Tarumi, Okayama University of Science

---

4:00 p.m. to 5:45 – The Best of Data Mining from KDD (Essex — Invited Session)
Session Organizer: Arnie Goodman, UCI
Session Chair: Arnie Goodman, UCI

4:00 p.m. — *Mining Concept-Drifting Data Streams Using Ensemble Classifiers*, Haixun Wang, IBM T. J. Watson Research Center

4:35 p.m. — *Mining Distance-Based Outliers in Near Linear Time*, Stephen Bay, Stanford University, and Mark Schwabacher, NASA Ames Research Center

5:10 p.m. — *Privacy Preserving K-Means Clustering over Vertically Partitioned Data*, Jaideep Vaidya, Purdue University, and Chris Clifton, Purdue University

---

4:00 p.m. to 5:45 – Computational Statistics I (Falkland — Contributed Session)

4:00 p.m. — *Estimating the Parameters of Infinite Scale Mixtures of Normals*, Hasan Hamdan, James Madison University, and John Nolan, American University (refereed)

4:20 p.m. — *Permutation Tests in Assessing Survival Forests for Prognosis Based on Gene Profiles*, Van L. Parsons, National Center for Health Statistics, and Thu M. Hoang, Universite Rene Descartes (refereed)

4:40 p.m. — *Wavelet Domain Linear Inversion via the LASSO*, Leming Qu, Boise State University, and Partha Routh, Boise State University (refereed)

5:00 p.m. — *Cramér-Rao Bounds and Monte Carlo Calculation of the Fisher Information Matrix in Difficult Problems*, James C. Spall, Johns Hopkins University, APL

5:20 p.m. — *User Profiling in Window Title and Process Table*, Chien-Chih Lin, School of Computational Sciences, George Mason Uniiversity, Eun Young Noh, School of Computational Sciences, George Mason University, Youngping Yan, School of Information Technology and Engineering, George Mason University, and Dr. Edward Wegman, School of Information Technology and Engineering, George Mason University

---

4:00 p.m. to 5:45 – Clustering (Galena — Contributed Session)
Chair: Jeff Solka

4:00 p.m. — *Cluster Subspace Identification Via Conditional Entropy Calculations*, James Diggans, George Mason University, and Jeffrey L. Solka, George Mason University

4:25 p.m. — *Actor Allegiance and Blockmodel Strength*, John Rigsby, Naval Surface Warfare Center, and Dr. Jeff Solka, Naval Surface Warfare Center

4:50 p.m. — *Bayesian Hierarchical Models of the Browsing Behavior of World Wide Web Users* , Juana Sanchez, University of California Los Angeles, and Ching-Ti Liu, University of California Los Angeles

5:15 p.m. — *Model-based Clustering With an Adaptive Mixtures Smart Start*, Jeffrey L. Solka, NSWCDD, and Wendy L. Martinez, ONR

7:00 p.m. to 10:00 p.m. – Banquet (Harborside Ballroom D/E)

Friday, May 28, 2004

7:30 a.m. — Registration Continues (Harborside Registration B)

8:00 a.m. to 9:45 a.m. – Best of the IASC I: High Dimensional Statistical Genomics from Genes to Proteins to Pathways (Harborside A — Invited Session)
Organizer: David Allison, University of Alabama at Birmingham
Chair: David Allison, University of Alabama at Birminghama

8:00 a.m. — *Using Bayesian Networks to Reconstruct Yeast Genetic Networks*, Grace S. Shieh, Institute of Statistical Science, Academia Sinica, Taiwan

8:35 a.m. — *Does Sequence Similarity Predict Expression Similarity*, Kui Zhang, Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, David Allison, Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Grier Page, Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, and Elliot J. Lefkowitz, Department of Microbiology, University of Alabama at Birmingham

9:10 a.m. — *Statistical Methods for Proteomics*, F. Seillier-Moiseiwitsch, University of Maryland - Baltimore County, Anindya Roy, University of Maryland - Blatimore County, and Yaming Hang, University of Maryalnd - Baltimore County

---

8:00 a.m. to 9:45 a.m. – Mass Spectroscopy and Clinical Proteomics (Harborside B — Invited Session) Organizer: Michael Trosset, William & Mary
Chair: Michael Trosset, William & Mary

8:00 a.m. — *Signal Conditioning and Filtering of SELDI Mass Spectrometry Time Series*, Dariya Malyarenko, Applied Science, College of William and Mary and INCOGEN, Inc., Dennis Manos, Applied Science, College of William and Mary, William Cooke, Physics, College of William and Mary, and Eugene Tracy, Physics, College of William and Mary

8:50 a.m. — *A Multiresolution View of Protein Mass Spectrometry Data*, Timothy W. Randolph, University of Washington, and Yutaka Yasui, Fred Hutchinson Cancer Research Center

---

8:00 a.m. to 9:45 a.m. – Text Mining and Applications (Essex — Invited Session)
Session Organizer: Ed Wegman, George Mason University
Session Chair: Ed Wegman, George Mason University

8:00 a.m. — *Identifying Cross Copora Document Associations Via Minimal Spanning Trees*, Jeffrey L. Solka, NSWCDD, Avory Bryant, NSWCDD, and Edward J. Wegman, George Mason University

8:35 a.m. — *Intersection Graphs for Text Analysis*, Elizabeth Leeds, NSWC, and David J. Marchette, NSWC

9:10 a.m. — *Document Classification and Clustering Using Weighted Text Proximity Matrices*, Wendy L. Martinez, Office of Naval Research, Angel R. Martinez, NSWCDD, and Edward J. Wegman, George Mason University

---

8:00 a.m. to 9:45 a.m. – Genetic Algorithms for Computational Biology (Falkland — Invited Session)
Organizer: John Grefenstette, George Mason University
Chair: John Grefenstette, George Mason University

8:00 a.m. — *Visual Data Mining of RNA Secondary Structure Folding Pathways as Determined by the Massively Parallel Genetic Algorithm*, Bruce A. Shapiro, Laboratory of Experimental and Computational Biology, NCI-Frederick, and Wojceich Kasprzak, Basic Research Program, SAIC-Frederick

8:35 a.m. — *Knowledge Discovery in Large Biological Data Sets Using Hybrid Classifier/Evolutionary Algorithms*, Michael Raymer, Wright State University, Mike Peterson, Wright State University, and Travis Doom, Wright State Unversity

9:10 a.m. — *Polyoptimizing Genetic Algorithms for Feature Selection*, Ewy Mathe, George Mason University, and John Grefenstette, George Mason University

---

8:00 a.m. to 9:45 a.m. – Classification (Galena — Contributed Session)

8:00 a.m. — *Comparison of Classification Techniques in Bioinformatics*, Rashpal Ahluwalia, West Virginia University, and Sundar Chidambaram, West Virginia University

8:25 a.m. — *Optimizing Bivalent Classifiers*, Jim DeLeo, National Institutes of Health Clinical Center

8:50 a.m. — *Confidence-Based Cost-Sensitive Classification Decisions*, Dragos D. Margineantu, The Boeing Company

9:15 a.m. — *A Two-Stage Nearest-Neighbor Classifier with Application to Microbial Source Tracking*, Jayson D. Wilbur, Department of Mathematical Sciences, Worcester Polytechnic Institute

---

9:45 a.m. to 10:30 a.m. – Coffee Break (Harborside Foyer)

10:30 a.m. to 12:15 p.m. – Best of the IASC II (Harborside A — Invited Session)
Organizer: Michael G. Schimek, University of Graz
Chair: Michael G. Schimek, University of Graz

10:30 a.m. — *Resampling Techniques in Neural Networks for Nonlinear time series analysis*, Michele La Rocca, Dept. of Economics and Statistics, University of Salerno, Italy, and Cira Perna, Dept. of Economics and Statistics, University of Salerno, Italy

11:20 a.m. — *Combining Ordinal Measures in Medical Research*, Knut M. Wittkowski, The Rockefeller University

---

10:30 a.m. to 12:15 p.m. – Statistical and Metrological Issues in Proteomics Using Time-of-flight Mass Spectrometry (Harborside B — Invited Session)
Organizer: Z. Q. John Lu, NIST
Chair: Z. Q. John Lu, NIST

10:30 a.m. — *Exploring Bioinformatics in Serum Proteomic Analysis for Early Detection of Prostate Cancer*, Bao-Ling Adam, Medical College of Georgia, Yutaka Yausi, Fred Hutchinson Cancer Research Center, Ziding Feng, Fred Hutchinson Research Center, and O. John Semmes, Eastern Virginia Medical School

10:55 a.m. — *Data-Driven and Peak-Based Feature Selection in Serum Protein Mass Spectrometry*, Walter S. Liggett, National Institute of Standards and Technology, Peter E. Barker, National Institute of Standards and Technology, O. John Semmes, Eastern Virginia Medical School, and Lisa H. Cazares, Eastern Virginia Medical School

11:20 a.m. — *Bioinformatics for Clinical Proteomics: Usage and Abusage*, Zhen Zhang, Johns Hopkins University, and Hong Zhang, Armstrong Atlantic State University

11:45 a.m. — *SVD-based Functional ANOVA For Measurement Evaluation of MALDI-TOF Mass Spectrometry*, Z. Q. John Lu, National Institute of Standards and Technology

---

10:30 a.m. to 12:15 p.m. – The Analysis of Streaming Data (Essex — Invited Session)
Organizer: Bill Szewczyk, NSA
Chair: Bill Szewczyk, NSA

10:30 a.m. — *Indexing Continual Range Queries for Efficient Stream Processing*, Kun-Lung Wu, IBM Watson Research, Shyh-Kwei Chen, IBM Watson Research, and Philip S. Yu, IBM Watson Research

11:05 a.m. — *Visual Analytics for Streaming Internet Data*, Ed Wegman, George Mason University and Karen Kafadar, (University of Colorado)

11:40 a.m. — *Streaming Graphics*, Leland Wilkinson, SPSS and Andrew Norton, SPSS

10:30 a.m. to 12:15 p.m. – Cancer Classification Using Gene Expression Profiling (Falkland — Invited Session)
Organizer: Dechang Chen, Uniformed Services University of the Health Sciences
Chair: Dechang Chen, Uniformed Services University of the Health Sciences

10:30 a.m. — *Cancer Prediction with Robust Kernel PLS Algorithm and Gene Expression Profile*, Zhenqiu Liu, Bioinformatic Cell/ TATRC, Dechang Chen, Uniformed Services University of the Health Sciences, and Jaques Reifman, Bioinformatic Cell/TATRC

11:05 a.m. — *Cancer Classification Using Informative Gene Profiles*, Xue-wen Chen, The University of Kansas

11:40 a.m. — *An Efficient Max-Dependency Algorithm For Gene Selection*, Hanchuan Peng, Lawrence Berkeley National Lab

---

10:30 a.m. to 12:15 p.m. – Trees (Galena — Contributed Session)

10:30 a.m. — *Some Statistical Issues Related to Feature Detection Using Random Forests*, Grant Izmirlian, National Cancer Institute (refereed)

10:55 a.m. — *Learning From Extremely Imbalanced Data with Random Forests*, Andy Liaw, Merck Research Labs, Chao Chen, UC Berkeley, and Leo Breiman, UC Berkeley

11:20 a.m. — *Tree Clustering of Gene Expression Data*, Lidia Rejto, University of Delaware, and Gabor Tusnady, Renyi Mathematical Institute, Budapest, Hungary

11:45 a.m. — *Some Light in a Dark Forest - a Closer Look at Tree Model Ensembles*, Simon Urbanek, University of Augsburg

---

12:15 p.m. to 1:45 p.m. — Lunch Break

1:45 p.m. to 3:30 p.m. – Highlights of the SAMSI Data Mining Year (Harborside A — Invited)
Organizer: David Banks, Duke
Chair: David Banks, Duke

1:45 p.m. — *Combinatorial Search in Data Mining*, David Banks, Duke University, and Leanna House, Duke University

2:20 p.m. — *Issues in 'Real Data' Mining*, Ashish Sanil, National Institute of Statistical Sciences

2:55 p.m. — *Mixtures of Factor Analyzers: Their Place in Data Mining*, Ernest Fokoue, SAMSI

1:45 p.m. to 3:30 p.m. – Epistasis (Harborside B — Invited)
Organizer: Bill Shannon, Washington University in St. Louis
Chair: Bill Shannon, Washington University in St. Louis

1:45 p.m. — *Gene-Gene and Gene-Environment Interactions in Genetic Case-Control Association Studies*, Jurg Ott, Rockefeller University, and Josephine Hoh, Yale University

2:20 p.m. — *Detecting Epistatic Interactions Contributing to Quantitative Traits*, Rob Culverhouse, Washington University School of Medicine, Tsvika Klein, Washington University School of Medicine, and William Shannon, Washington University School of Medicine

2:55 p.m. — *Systems Biology Thought Experiments for Interpreting Epistasis Models*, Jason Moore, Vanderbilt University

---

1:45 p.m. to 3:30 p.m. – Analysis of Very Large Data Sets (Essex — Invited)
Organizer: David Scott, Rice University
Chair: David Scott, Rice University

1:45 p.m. — *Having It All*, Allan Wilks, AT&T Labs - Research

2:20 p.m. — *Interactive Graphics for Large Data Sets: There is More to it Than Meets the Eye*, Antony Unwin, Augsburg University

2:55 p.m. — *Alternatives to Mixture Modeling in High Dimensions*, David Scott, Rice University

---

1:45 p.m. to 3:30 p.m. – Genetic and Biochemical Networks: Methods and Empirical Models (Falkland — Invited)
Organizer: Nicholas J. I. Lewin-Koh, Lilly Systems Biology Pte Ltd
Chair: Nicholas J. I. Lewin-Koh, Lilly Systems Biology Pte Ltd

1:45 p.m. — *Limitations of Statistical Learning from Gene Expression Data*, Tianjiao Chu, Institute for Human and Machine Cognition, University of West Florida, and Clark Glymour, Institute for Human and Machine Cognition, University of West Florida

2:20 p.m. — *Generating Constraints on the Topology of Genetic Networks Using Expression Data: a Combinatorial Approach*, Nicholas J. I. Lewin-Koh, Lilly Systems Biology Pte Ltd, and Christopher Taylor, Lilly Systems Biology Pte Ltd

2:55 p.m. — *Transient Response of Steady-State Yeast Cells to Small Perturbations*, Michal Ronen, Stanford, and David Botstein, Princeton

---

3:30 p.m. to 4:00 p.m. – Refreshment Break (Harborside Foyer)

4:00 p.m. to 5:45 p.m. – Visualization and Analysis of Text/Web Data (Harborside A — Invited)
Organizer: Dunja Mladenic, J. Stefan Institute
Chair: Dunja Mladenic, J. Stefan Institute

4:00 p.m. — *Efficient Visualization of Large Text Corpora*, Marko Grobelnik, J.Stefan Institute, and Dunja Mladenic, J.Stefan Institute

4:50 p.m. — *Visual Text Mining with TRUST and Starlight*, Anne Kao, Boeing, John Risch, Pacific Northwest National Lab, Steve Poteet, Boeing, and Jason Wu, Boeing

---

4:00 p.m. to 5:45 p.m. – Protein Folding (Harborside B — Invited)
Organizer: David Banks, Duke
Chair: David Banks, Duke

4:00 p.m. — *Statistical and Computational Issues in AB Initio Protein Structure Prediction*, Ingo Ruczinski, Johns Hopkins University

4:35 p.m. — *Finding the Protein-Protein Interface via Docking Calculations*, Jeffrey J. Gray, Chemical & Biomolecular Engineering,Johns Hopkins University

5:10 p.m. — *Five Hierarchical Levels of Sequence-Structure Correlation in Proteins*, Chris Bystroff, Department of Biology, Rensselaer Polytechnic Institute

---

4:00 p.m. to 5:45 p.m. – West Nile (Essex — Invited)
Organizer: Juergen Symanzik, Utah State University
Chair: Juergen Symanzik, Utah State University

4:00 p.m. — *The Anatomy of a Bioevent: West Nile Virus in Washington, DC*, James Wilson, ISIS Center, Georgetown University

4:35 p.m. — *West Nile Virus, Dead Birds and Human Infections. A Geostatistical Approach.*, Sean C. Ahearn, Center for Advanced Reesearch of Spatial Information (CARSI Lab.), Hunter College-CUNY, Constandinos Theophilides, Center for Advanced Research of Spatial Information (CARSI Lab.), Hunter College-CUNY, and Edward S. Binkowski, Department of Mathematics and Statistics, Hunter College-CUNY

5:10 p.m. — *Visualization, Web-Access, and Simulation of West Nile Virus Data - From the Regional to the National Level*, Juergen Symanzik, Utah State University, Robert Gillies, Utah State University, Samson Gebraeb, Utah State University, Gopi Krishna, Utah State University, Peter Ma, Utah State university and James Wilson, ISIS Center, Georgetown University

---

4:00 p.m. to 5:45 p.m. – Statistical Analysis of Internet Data (Falkland — Invited)
Organizer: Al Hero, University of Michigan
Chair: Al Hero, University of Michigan

4:00 p.m. — *Network Tomography*, Bin Yu, UC Berkeley

4:35 p.m. — *Empirical Analysis of Structure in Computer Network Traffic Flows*, Eric D. Kolaczyk, Boston University

5:10 p.m. — *Hierarchical Clustering and Network Topology Identification*, Rui Castro, Rice University, Robert Nowak, University of Wisconsin, Madison, and Mark Coates, McGill University, Canada

---

Saturday, May 29, 2004

7:30 a.m. — Registration Continues (Harborside Registration B)

8:00 a.m. to 9:45 a.m. – Mixture Modeling of Gene Expression Data (Harborside A — Invited Session)
Organizer: Kim-Anh Do, George Mason University
Chair: Kim-Anh Do, George Mason University

8:00 a.m. — *Mixture Models in Molecular Classification*, Giovanni Parmigiani, Johns Hopkins University

8:35 a.m. — *Supervised and Unsupervised Learning Methods for Gene-Expression Data*, Geoff McLachlan, University of Queensland, and Christophe Ambroise, University of Compiegne

9:10 a.m. — *Clustering-Based Classification for Gene Function Prediction Using Microarray Data*, Wei Pan, University of Minnesota, and Guanghua Xiao, University of Minnesota

---

8:00 a.m. to 9:45 a.m. – Bioinformatics I (Essex — Contributed Session)

8:00 a.m. — *A Novel Method for Estimating Scale Recombination Rate from Sequence Data*, Osho O. Ajayi, University of Reading, England, UK

8:25 a.m. — *An Efficient Algorithm for Simulating Coalescence with Recombination.*, Katy L. Simonsen, Statistics Department, Purdue University, Dan A. Noland, CS Department and ITaP, Purdue University, and Chinh Le, ITaP, Purdue University

8:50 a.m. — *Multi-protein Complex Data Clustering for Detecting Protein Interactions and Functional Organizations*, Chris Ding, Lawrence Berkeley National Laboratory, Xiaofeng He, Lawrence Berkeley National Laboratory, Richard Meraz, Lawrence Berkeley National Laboratory, and Steve Holbrook, Lawrence Berkeley National Laboratory

9:15 a.m. — *DNAMR and DNAMRweb, Developing Easy to Use Software for DNA Microarray Data Mining*, Vladimir Kovtun and Cabrera Amaratunga, Rutgers

---

8:00 a.m. to 9:45 a.m. – Computational Biology (Falkland — Contributed Session)

8:00 a.m. — *Probabilistic Classification in High Dimensions for Drug Discovery*, Alexander Gray, Carnegie Mellon University, Paul Komarek, Carnegie Mellon University, Ting Liu, Carnegie Mellon University, and Andrew Moore, Carnegie Mellon University

8:25 a.m. — *Computation of the kth Nearest Neighbor Estimate of Entropy of Molecules Using Parallel Processing*, E. James Harner, West Virginia University, Jun Tan, West Virginia University, Shengqiao Li, West Virginia University and NIOSH, and Harshinder Singh, West Virginia University and NIOSH

8:50 a.m. — *Assessment of the Relative Therapeutic Effect in Small Groups at Several Time Points: Efficacy of Mucosal and Subcutaneous Peptide Vaccines IN Rhesus Macaques Exposed to SHIV* , Vladimir A. Kuznetsov, SRA International, Inc.&CIT/NIH, Bethesda MD, Vladimr S. Stepanov, Central Economic and Mathematics Institute, RAS, Moscow, Russia, Jay A. Berzofsky, NCI/NIH, Bethesda MD, and Igor M. Belyakov, NCI/NIH, Bethesda MD

9:15 a.m. — *Identifying Differentially Expressed Proteins in 2-D DIGE Experiments*, Yan Ma, West Virginia University, and E. James Harner, West Virginia University

---

8:00 a.m. to 9:45 a.m. – Large Datasets (Falkland — Contributed Session)

8:00 a.m. — *Single-pass, Low-storage Methods for Massive Streaming Datasets with Applications to Multivariate Density Estimation*, James P. McDermott, Bristol-Myers Squibb, and Dennis K. J. Lin, Pennsylvania State University

8:25 a.m. — *Fitting Large-Scale Spatial Models with Applications to Microarray Data Analysis*, Stephan R. Sain, CU-Denver, and Reinhard Furrer, NCAR

8:50 a.m. — *Parallelizing the Computation of Spatial Covariance in Large Spatial Data Sets*, James A. Shine, US Army Topographic Engineering Center

9:15 a.m. — *Computationally Efficient Identification of Outliers in Large Data Sets*, Mark Werner, Oakland University

---

9:45 a.m. to 10:30 a.m. – Refreshment Break (Harborside Foyer)

10:30 a.m. to 12:15 p.m. – Towards Understanding and Analyzing Proteomics Data (Harborside A — Invited Session)
Organizer: Kim-Anh Do, George Mason University
Chair: Kim-Anh Do, George Mason University

10:30 a.m. — *Bayesian Methods for Proteomics with Feature Selection*, Marina Vannucci, Texas A&M University, Mahlet G. Tadesse, Texas A&M University, and Jeffrey S. Morris, MD Anderson Cancer Center

11:05 a.m. — *The Analysis of MALDI-TOF Proteomic Spectra from Serum Samples - A Case Study*, Keith Baggerly, MD Anderson Cancer Center

11:40 a.m. — *Nonparametric Approaches to the Classification of Proteomic Profiles*, Kim-Anh Do, U. T. M.D. Anderson Cancer Center, Peter Mueller, U. T. M. D. Anderson Cancer Center, Sijin Wen, U. T. M. D. Anderson Cancer Center, and Raj Bandyopadhya, Rice University

10:30 a.m. to 12:15 p.m. – Bioinformatics II (Essex — Contributed Session)

10:30 a.m. — *Genome Phylogenetic Analysis Based on Extended Gene Contents*, Hongmei Zhang, University of West Flordia, and Xun Gu, Iowa State University

10:55 a.m. — *Characterization and Re-Annotation of Common Genes Found in Complete Chloroplast Genomes*, Beatrice Kilel, George Mason University

11:20 a.m. — *Data Analysis and Modeling of the Evolution of Proteome Complexity*, Vladimir A. Kuznetsov, SRA International, Inc.&CCIT/NIH, Bethesda MD

11:45 a.m. — *Modeling Dinucleotide Density Fluctuations in Genome Sequences*, R. H. Baran, Office of Naval Research and Naval Surface Warfare Center

---

10:30 a.m. to 12:15 p.m. – Matrix Computations and Data Mining (Falkland — Invited Session)
Organizer: Jesse Barlow, Pennsylvania State University
Chair: Jesse Barlow, Pennsylvania State University

10:30 a.m. — *Principal Component and Self-aggregation Clustering*, Chris Ding, Lawrence Berkeley National Laboratory

10:55 a.m. — *Operations to Construct and Maintain a Truncated ULV Decompostion*, Jesse Barlow, The Pennsylvania State University

11:20 a.m. — *Classification of Microarray Data by Two-way Gaussian Mixtures*, Jia Li, The Pennsylvania State University

11:45 a.m. — *Unified Multiclass Proximal Support Vector Machines*, Hao Helen Zhang, North Carolina State University

---

10:30 a.m. to 12:15 p.m. – Computational Statistics II (Galena — Contributed Session)

10:30 a.m. — *On Two Sample Data Analysis*, Sujung Choi, Texas A&M University

10:50 a.m. — *Jointly Optimizing Model Complexity and Data-Processing Parameters*, Jim Garrett, PhD, Becton Dickinson (refereed)

11:10 a.m. — *Noncentral Generalized F Distributions with Applications to Joint Outlier Detection*, Donald E. Ramirez, University oof Virginia (refereed)

11:30 a.m. — *Performance Metrics for Group-Detection Algorithms*, James V. White, Alphatech, Inc., Sam Steingold, Alphatech, Inc., and Connie G. Fournelle, Alphatech, Inc. (refereed)

11:50 a.m. — *Monte Carlo Analysis of Univariate Robust Statistical Outlier Techniques*, Mark W. Lukens, George Mason University, and James E. Gentle, George Mason University

Interface 2004 Abstracts

Thursday, 8:45 a.m. to 9:45 a.m. – Keynote Address

Computational Biology and the Ribosome

David Draper (Johns Hopkins University), draper@jhu.edu

Abstract

The translation of genetic information into functional proteins is mediated by the ribosome, a ubiquitous and highly conserved cellular organelle. The very large size of the ribosome, the complexity of its composite protein and RNA structure, and the number of its interactions with cellular factors have made elucidation of its functional mechanism an extremely challenging problem. Remarkable progress has been made in recent years, to the point where many questions about ribosome structure, function, and evolution can be formulated in precise molecular terms. Bioinformatics and computational biophysics have been important contributors in these advances. This talk will give an overview of the computational challenges posed by ribosome studies, with reference to the ways sequence databases, structure-based calculations, and other computations have helped guide ribosome studies on specific topics ranging from fundamental questions in RNA folding to design of high-throughput screens for antibiotics.

Thursday, 10:30 a.m. to 12:15 a.m. – Tubes and Graphs (Invited Session)

The Volume-of-Tube formula: Computational Methods and Statistical Applications

Catherine Loader, (Case Western Reserve University), catherine@cwru.edu, and Ramani S. Pilla, (Case Western Reserve University), pilla@cwru.edu

Abstract

The Volume-of-Tube formula, originally developed by Hotelling (1939 Amer. J. Math. 440-460) for curves and extended by Weyl (1939 Amer. J. Math. 461-471) for surfaces, evaluates the volume of neighborhoods manifolds lying in Euclidean space or on the surface of spheres. In statistics, the results can be used to approximate probabilities arising from the maximum of random fields. Applications include testing for the presence of a nonlinear term in regression models; construction of simultaneous confidence bands around regression surfaces; functional analysis-of-variance problems; detection of nonuniformity in spatial scan analysis and testing for the number of components in mixture models.

Each of these problems define a corresponding manifold lying on the (possibly infinite dimensional) unit sphere. Implementation of the tube formula requires determination of the geometry of the manifold and in particular its boundary behavior. In this talk, we discuss computational procedures for determining the geometric constants appearing in the Hotelling-Weyl results and in boundary corrections provided by Naiman (1990 Ann. Statist. 685-716). In applications, the manifold may be defined either through an explicit vector representation, or through a distance metric (or covariance) function, and we discuss how to implement the formula in each case. A software library, Libtube, that implements the Hotelling-Weyl-Naiman results up to fourth order terms, will be presented. We also discuss the relative merits of the tube formula versus simulation and other methods for approximating supremum distributions.

Multivariate Statistics for Trees and Networks

Susan Holmes, (Stanford University), susan@stat.stanford.edu

Abstract

We will show examples of analysing data from various sources that produce trees and networks that we want to compare. These are enriched graphs with variables at their nodes and sometimes also on their edges. We define useful distances on both trees and networks and then project the data into a multidimensional euclidean space that enables effective pattern searching. Examples will be drawn from work on comparing phylogenetic trees on microbial systems, birds comparing hierarchical clustering trees on microarray data and comparing metabolic networks.

Thursday, 10:30 a.m. to 12:15 a.m. – Internet Tomography (Invited Session)

Wavelet and SiZer Analyses of Internet Traffic Data

Cheolwoo Park, (SAMSI), cwpark@samsi.info

Abstract

It is important to characterize burstiness of Internet traffic and find the causes for building models that can mimic real traffic. To achieve this goal, exploratory analysis tools and statistical tests are needed, along with new models for aggregated traffic. This talk introduces statistical tools based on wavelets and SiZer (SIgnificance of ZERo crossings of the derivative). The intricate fluctuations of Internet traffic are explored in various respects and lessons from real data analyses are summarized.

Tools for the Analysis and Visualization of Network Traffic

George Michailidis, (The University of Michigan), gmichail@umich.edu

Abstract

In this talk we provide a framework for analyzing network traffic traces through trace-driven queueing. We also introduce several queueing metrics together with the associated visualization tools that provide insight into the traffic features and facilitate comparisons between traces. Some techniques for non-stationary data are discussed. Applying our framework to both real and synthetic traces we illustrate how to compare traces using trace-driven queueing, and also show that traces that look 'similar' under various statistical measures (such as the Hurst index) can exhibit rather different behavior under queueing simulation.

Thursday, 10:30 a.m. to 12:15 a.m. – Too Many Features, Too Few Samples – Challenges for Disease Profiling Using Biomedical Data (Invited Session)

The Analysis of Biomedical Data - Caveats and Challenges

R.L. Somorjai, (Institute for Biodiagnostics, National Research Council Canada), ray.somorjai@nrc-cnrc.gc.ca

Abstract

Biomedical data- magnetic resonance, infrared, fluorescence, Raman spectra of biofluids and tissues, as well as microarray expression profiles from genomics and mass spectra from proteomics are characterised by a large number (O(1000 - 10000) of features (genes, M/Z values, spectral frequencies) and relatively few (O(10- 100) samples. This leads to the twin curses of dimensionality and dataset sparsity. The talk will focus on the statistical consequences of these curses when the goal is the development of robust, reliable classifiers. Some approaches to lift the curses will be discussed, and the major remaining challenges and possible solutions outlined. A simple approach for the visualization of high-dimensional data will also be demonstrated.

Supervised Analysis When the Number of Features (p) Greatly Exceeds the Number of Cases (n)

Richard Simon, (National Cancer Institute), rsimon@nih.gov

Abstract

New genomic and proteomic technologies provide measurements of thousands of features for each case assayed. This provides a context for enhanced discovery and false discovery. Most statistical procedures were not developed for the p ¿¿ n setting and the literature of DNA microarray studies contains many examples of misuse of analytic procedures such as cross-validation. This paper will highlight some of the most serious potential problems. We will discuss misconceptions concerning the application of sample re-use methods such as cross-validation and the bootstrap for class prediction problems. We also describe the use of multivariate permutation test for class comparison problems in the p ¿¿ n setting.

Many features, few samples: from cheminformatics to bioinformatics

Kristin P. Bennett, (Rennselaer Poly. Tech.), bennek@rpi.edu,
Curt Breneman, (Rensselaer Poly. Tech), brenec@rpi.edu, and
Mark Embrechts, (Rensselaer Poly. Tech.), embrem@rpi.edu

Abstract

Many problems in bioinformatics and cheminformatics have similar characteristics – many highly-correlated features with few data points. We examine methodologies developed for predicting bioactivities of small molecules to support drug discovery and show how they can be applied to problems in bioinformatics. Highly regularized kernel methods such as support vector machines and kernel partial least squares regression combined with feature selection and variance reduction techniques are used both for prediction and to provide insights into the underlying mechanisms.

Thursday, 10:30 a.m. to 12:15 a.m. – Visualization (Contributed Session)

Extensions to Dynamically Conditioned Choropleth Maps

Yuguang Zhang, (FreddieMac), yuguang_zhang@freddiemac.com , and
Daniel B. Carr, (George Mason University), dcarr@gmu.edu

Abstract

This paper discusses recent extensions of CCmaps. CCmaps is visual analytic software for dynamically conditioning and studying choropleth maps. The software serves as a hypothesis generation tool for patterns observed in geospatially indexed statistics. Since CCmaps incorporates many basic statistics and statistical graphics, such as weighted means, two-way tables, scatterplot smoothes and QQplots, CCmaps can also

motivate discussion in statistical education settings. The first extension described is a fast search algorithm for creating a pick list of interesting slider settings. The second is the ability to save annotated sequences of the program state for later replay. CCmaps is fully live when each state is restored. This fundamental analysis management feature for dynamic software is also useful for presentations and in the production of tutorials.

Visual Analytics for Dynamically Conditioned Choropleth Maps: QQplots, Scatterplot Smoothes and Two-Way Tables

Chunling Zhang, (George Mason University), czhang1@gmu.edu,
Yaru Li, (George Mason University), yli6@gmu.edu, and
Daniel Carr, (George Mason University), dcarr@gmu.edu

Abstract

The visual analytics presented in this paper augment conditioned choropleth maps. In the conditioned map, dynamic sliders partition the map into a 3 x 3 grid of partial maps. Two different variables are attached to the two partitioning sliders. One slider controls row membership in the grid and their other controls column membership. The analysts visual impression and comparison of the partial maps can be made more quantitative by showing other analytics. The analytics described in this talk are modifications of conventional QQplots, smoothed scatterplots, and two-way tables of means, effects, and model statistics. One modification involves the use of weights. Most modifications speed the response in order to keep up with the dynamic partitioning sliders. For example the smoothing widgets include the option to use an intermediate binning step when thousands of regions are involved. The talk provides live examples. The applications involve different kinds of region such as county elementary school districts, hexagon grids for three states, and nations of the world.

Widget Management in the Multilayered 3-D Visual Analytic Software Called Glisten

Yanling Liu, (George Mason University), yliu6@gmu.edu, and
Daniel Carr, (George Mason University), dcarr@gmu.edu

Abstract

This paper describes a flexible approach to managing widgets. The need for widget management arises in multilayered 3-D software called Glisten. Glisten is typically used for rendering glyphs that encode statistics. Color control, size control and multiple filtering widgets provide dynamic control of the transformation from variables into properties of 2-D point or 3-D sphere glyphs. More complex glyphs and multiple layers of glyphs used to represent many variables can result in widgets proliferation.

Widget management allows the analyst to easily select, hide, and rearrange widgets to help focus on active widgets and minimize distraction. One live example will show rendering control for statistics associated with peptide docking on immune system molecules. Another will show a 3-D protein backbone and statistics related to protein tessellating tetrahedra indexed by amino acids. Glisten with widgets, rotation, zoom and other features provides an interesting environment for rendering protein features and statistics in the same 3-D view.

Visualizing Patients Treated with Three-Dimensional Computed Tomography-Guided Brachytherapy

Faleh Alshameri, (George Mason University), falshame@gmu.edu
Jee Vang, (George Mason University), jvang@scs.gmu.edu

Abstract

This paper focuses on visual data mining methods for prostate cancers. The most common cancer among US men is prostate cancer. It is estimated that in the year 2002, 189,000 US men were diagnosed with prostate cancer, of which 30,200 died. Even with these alarming numbers, there is still no universally agreed-upon strategic plan for the diagnosis and management of prostate cancer. One of the contributing factors to why there is no universal treatment of prostate cancer is because treatment requires individualized care. Depending on the details of the prostate cancer, such as severity, locality, recurrence, patient age, patient life expectancy, etc..., a particular treatment method or cocktail of management care is prescribed. In this paper, we explore these variables using the CrystalVision data mining tool.

Multidimensional Scaling and Classification

Michael W. Trosset, (College of William & Mary), trosset@math.wm.edu

Abstract

Most classification methods assume that the objects to be classified have been embedded in a Euclidean space. If another measure of dissimilarity is preferred, then the objects can be embedded in a Euclidean space via multidimensional scaling (MDS). The use of MDS to preprocess data for classification poses various conceptual and technical challenges. Traditional MDS is unsupervised, more commonly used for clustering than for classification. Because traditional MDS maps a finite set of objects into Euclidean space, it is not clear how to construct a classifier that can be applied to new cases. Novel formulations of MDS address these challenges.

Thursday, 10:30 a.m. to 12:15 a.m. – Microarrays I (Contributed Session)

Extending the Loop Design for Microarray Experiments

Naomi S. Altman, (Pennsylvania State University), naomi@stat.psu.edu

Abstract

The loop design (Kerr and Churchill, 2001) is a clever application of incomplete blocks of size 2 to 2-channel microarray experiments. In this talk, I discuss the application of the loop design to experiments with multiple factors, blocks and other situations, and discuss efficiency and robustness to missing data.

Mixed Effects Model for Assessing RNA Degradation in Affymetrix GeneChip Experiments

Kellie J. Archer, Ph.D., (Department of Biostatistics, Center for the Study of Biological Complexity, Virginia Commonwealth University), kjarcher@vcu.edu,
Suresh E. Joel, (Department of Biomedical Engineering, Virginia Commonwealth University), sejoel@vcu.edu, and
Viswanathan Ramakrishnan, (Department of Biostatistics, Virginia Commonwealth University), vramesh@mail2.vcu.edu

Abstract

Due to the high cost of microarray experiments, investigators typically select designs with biological rather than technical replicates. Therefore, it is essential that the quality of RNA hybridized to the microarray meets certain standards. The process of transcription begins with reverse transcriptase binding at the 3' end of a gene and continuing toward the 5' end. However, transcription generally does not continue to completion. That is, reverse transcription typically drops off before reaching the 5' end. Affymetrix GeneChips includes probe sets which interrogate both the 3' end and the 5' end for selected control genes to assess quality of transcription. The MAS 5.0 software estimates the 3'/5' ratio after the PM and MM probes have been summarized into a probe set expression measure. Unfortunately, inherent to all probe set expression summary methods is that the 3' and 5' probe sets of interest are only represented on the GeneChip once. This leads to the unfortunate consequence of inadequate replications for variance calculation. The methodology proposed uses the pixel level intensities to increase the number of observations per probe set in order to obtain a better estimate of the 3' to 5' ratio. Since there is an inherent hierarchical structure to GeneChip data, where pixels are nested within probes and probes are nested within probe sets, RNA degradation will be assessed by fitting mixed effects ANOVA models to estimate the 3' to 5' ratio treating probe set as a fixed effect, while treating pixel level and

PM level data as random effects. This enables the construction of confidence intervals about the estimated ratio. The estimated confidence interval will more appropriately indicate whether the RNA was of sufficient quality rather than judging quality based on the ratio being below an arbitrarily selected threshold. Results from HG-U133A GeneChips will be presented.

Performance of the False Discovery Rate for Small Sets of cDNA Microarrays

Simon Rosenfeld, (National Cancer Institute), sr212a@nih.gov

Abstract

Detection of differentially expressed genes from microarray data is particularly difficult when the number of replicates available for the analysis is small (what usually takes place in practice) and/or genomic signal is weak. Using Monte Carlo simulation, we analyze the performance of a popular approach for detecting differentially expressed genes based on application of the Benjamini- Hochberg False Discovery Rate. We show that this performance, measured by its ability to detect differentially expressed genes, is fairly weak when the number of replicates is small and/or genomic signal is weak. The reason for this is rooted in the very nature of the approach, which associates the statistical significance with the smallest p-values. We suggest a methodology for increasing the sensitivity of gene discovery based on a new concept of bio-weight. We also discuss related methodological topics, such as applicability of the t-test and permutation test for small non-normal samples.

Mutual Information Based ICA Applied to Microarray Data

Hoang, M. Thu, (Universite Rene Descartes), hoang@biomedicale.univ-paris5.fr,
Pham, Dinh Tuan, (Centre National de la Recherche Scientifique, IMAG),
Dinh-Tuan.Pham@imag.fr, and
Thierry-Pascal Baum, (Centre National de la Recherche Scientifique, IMAG),
Thierry-Pascal.Baum@imag.fr

Abstract

In this contribution, we demonstrate the use of the latest and performant algorithm for independent component analysis (ICA) developed by one of the author [1] for analyzing microarray data. We illustrate on T-ALL microarray data also discussed in [2] as an example in using gene profiling for prognosis. We intend to compare our method to similar statistical methods for relevance in extracting clinical knowledge from gene expression in human leukemia.

[1] D. T. Pham (2004) Fast algorithms for Mutual Information Based Independent Component Analysis. IEEE Trans. on Signal Processing. To appear. (http://www-lmc.imag.fr/lmc-sms/Dinh-Tuan.Pham/BSS/AlgoICAinfof.ps.gz)

[2] Hoàng T. and Parsons V. (2004) Bagging Survival Trees for Prognosis based on Gene Profiles, Compstat'2004. To appear. Physica Verlag.

Thursday, 1:45 p.m. to 3:30 – Mixture Modeling: Modern Approaches and Applications (Invited Session)

The Volume-of-Tube Formula: Applications to Perturbation and Mixture Models

Ramani S. Pilla, (Case Western Reserve University), pilla@case.edu, and
Catherine Loader, (Case Western Reserve University), catherine@cwru.edu

Abstract

In this talk, a general class of models called "perturbation models" are introduced. These models are described by an underlying "null" model that accounts for most of the structure in a data while a perturbation accounts for possible small localized departures. The theory and inferential methods for fitting the perturbation models are discussed. Two important statistical problems for which perturbation models are applicable are finite mixture models and spatial scan analysis. In a finite mixture model, the null model represents a mixture with m number of components and the perturbation model represents additional components. In the spatial scan analysis context, the null density models the background or noise whereas the perturbation searches for an unusual region such as a tumorous tissue in mammography or a target in automatic target recognition.

The theory based on the Hotelling-Weyl tube formula provides an elegant approach to solving classes of statistical problems for which exact solution is intractable and when regularity conditions required for the application of the central limit theorem type results are not satisfied. In this talk, (1) a new test statistic for detecting the presence of perturbation is described, (2) a general theory for the derivation of the limiting distribution of the test statistic is illustrated using the results of Hotelling-Weyl tube formula, (3) computational issues associated with the implementation of the procedures are addressed and (4) the resulting theory is applied to the problem of testing for the number of terms in mixture models. Application of the resulting theory to image analysis will be discussed.

The Role of Latent Variables in Model Selection Accuracy

Jeremy Nadolski, (University of Kentucky), jeremy@ms.uky.edu

Abstract

Mixture models are often formulated in terms of latent variables Z which determine the component membership of the observations. While these latent variables are often used solely as a computational tool, we will discuss how the latent variable formulation provides insight into model selection procedures. We will demonstrate conditions on the latent variables that cause BIC (and other model selection procedures) to fail, and suggest alternative methods for model selection more suited for those conditions.

Empirical Likelihood Based Inferences in Semiparametric Finite Mixtures

Jing Qin, (Memorial Sloan-Kettering Cancer Center), qinj@biost.mskcc.org

Abstract

Marginal likelihood and conditional likelihood are the most popular methods to eliminate nuisance parameters. For a parametric model, it is well known that the full likelihood can be decomposed into the product of a conditional likelihood and a marginal likelihood. This property is less transparent in nonparametric or semiparametric likelihood setting. In this talk we show that this nice parametric likelihood property can be carried over to the empirical likelihood (Owen (1988)) world. More importantly, we find some of its applications in case-control studies, genetical linkage analysis, genetical quantitative trials analysis, tuberculosis infection data, and unordered pairs. All these applications can be treated as semiparametric finite mixture models. An analogous version of EM algorithm in parametric setup is discussed under a semiparametric density ratio model.

Thursday, 1:45 p.m. to 3:30 – Analysis of Functional Neuroimaging Data (Invited Session)

Brains on Film: Using Optical Imaging to Build Maps of Brain Activity

Kary Myers, (Carnegie Mellon University), kary@stat.cmu.edu

Abstract

Through functional neuroimaging techniques, scientists can construct maps of the brain to indicate regions of stimulus-induced activity. I will present maps constructed with

data from optical imaging experiments - experiments in which a high resolution video camera records a 'movie' of the surface of the active brain. In my data, each pixel is roughly 12 microns on a side,, each video frame has 480 x 640 pixels, and we record 30 frames per second. While this movie of activation has better spatial resolution than images from any other in-vivo method, the activation is masked by a tremendous amount of noise, with a typical signal-to-noise ratio of 0.001. Because the noise primarily comes from physiological fluctuations that are unrelated to the experimental stimulus, I reduce this noise by taking advantage of simultaneously recorded physiological data like blood pressure and EEG, thus building better maps of the brain.

Reducing Physiological Noise in fMRI Experiments

Rebecca L. McNamee, (University of Pittsburgh), rlandes@stat.cmu.edu

Abstract

The nearly periodic oscillations of heart beat and respiration are a prominent source of data contamination in functional Magnetic Resonance Imaging (fMRI) studies. Physiological noise can lead to complex distortions in the data, which can in turn affect the quality of the results. Methods for reducing this type of noise have been previously examined by several investigators. We present a highly effective yet simple method for routine use in reducing the effects of physiological noise. The technique uses linear regression while allowing for time delays between the recorded physiological data and its effect on the fMRI signals. This time delay was found to be important in the case of the cardiac data but not in the case of the respiration.

Wavelet-Based Statistical Analysis of fMRI Data

Ivo Dinov, (UCLA Statistics/Neurology), dinov@stat.ucla.edu,
Arthur Toga, (UCLA Neurology), toga@loni.ucla.edu,
Michael Mega, (Neural Net Research), mega@loni.ucla.edu, and
John Boscardin, (UCLA Biostatistics), jbosco@ucla.edu

Abstract

We propose a new method for statistical analysis of functional magnetic resonance imaging (fMRI) data. The discrete wavelet transformation is employed as a tool for efficient and robust signal representation. We use structural MRI and functional fMRI to empirically estimate the distribution of the wavelet coefficients of the data both across individuals and across spatial locations. Heavy-tail distributions are then proposed to model these data because these signals exhibit slower tail decay than the Gaussian distribution. There are two basic directions we investigate in the first part of this study: 1. Bayesian wavelet-based thresholding scheme, which allows better signal representation, and; 2. A family of heavy-tail distributions, which are used as models for the

real MRI and fMRI timeseries data. We discovered that Cauchy, Bessel K-Forms and Pareto distributions provide the most accurate asymptotic models for the distribution of the wavelet coefficients of the data. In the second part of our investigation we will apply this technique to analyze a large fMRI data set involving repeated presentation of sensory-motor response stimuli in young, elderly and demented subjects.

Thursday, 1:45 p.m. to 3:30 p.m. – Text Mining for Biomedical Applications (Invited Session)

Natural Language Processing for Biosurveillance

Dr. Wendy W. Chapman, (University of Pittsburgh), chapman@cbmi.pitt.edu

Abstract

Much of the pre-existing electronic data that could be harnessed for early outbreak detection is in free-text format. Natural language processing (NLP) techniques may be useful to biosurveillance by classifying and extracting information described in free-text sources. In the Real-time Outbreak and Disease Surveillance laboratory we are developing and evaluating NLP techniques for surveillance of syndromic presentations and specific findings or diseases of interest in public health.

I will describe the NLP techniques we have applied to dictated clinical reports for this task. I will also describe the three levels of evaluation we believe are crucial to understanding whether NLP techniques contribute to outbreak detection. First, I will describe the performance of various feature detection tasks in which we extract or classify information from clinical text. Second, I will describe how well the extracted features detect the cases of interest in biosurveillance. Third, I will describe initial work showing a correlation between classification of patients from free-text and disease outbreaks.

Automated Terminological Networks For High Throughput Comparative Biology of Phenotypes

Yves A. Lussier, (Columbia University), lussier@dbmi.columbia.edu, and
Xiaoyan Wang, (Columbia University), xw108@columbia.edu

Abstract

The integration of heterogeneous phenotypic databases is increasingly important as cross-disciplinary research escalates. However, manually indexing of every phenotype in large-scale databases is laborious, remarkably time consuming and often impractical. In addition, non-molecular phenotypes are complicated: phenotypic scales span

from molecules to the organism, their scopes are somewhat undetermined and their (non-molecular) representation standards are generally not expressive enough to automatically annotate radically new phenotypes. This study has examined the systematic use of terminology and knowledge based technologies to enable high-throughput comparative phenomics. We have developed a system (Clinigene) that uses multistrategy approaches to clarify and relate phenotypes according to their semantic, probabilistic and terminological properties. In a cluster of databases and terminologies consisting of OMIM, SNOMED 3.5, SNOMED CT, UMLS and Valle's human gene-disease database, we have demonstrated that the can increase the precision and recall over current rate-limiting mapping approaches, such as manual curation. In addition, Clinigene's Automated Terminological Networks (ATN) bridged the phenotype gap between phenotypic databases containing related phenotypes expressed in different granularities.

The MiTAP System for Monitoring Reports of Disease Outbreak

Laurie Damianos, (The MITRE Corporation), laurie@mitre.org

Abstract

The MiTAP system was developed as an experimental prototype using human language technologies for monitoring infectious disease outbreaks and other global disasters. MiTAP is designed to provide timely multi-lingual information access to analysts, medical experts, health services, and individuals involved in humanitarian assistance and relief work. Every day, thousands of articles from hundreds of information sources are automatically captured, filtered, translated, tagged, summarized, categorized by content, and made available to users via a news server and web-based search engine. Information extraction technology plays a critical role in many of these processes, presenting information in a variety of time-saving mechanisms to facilitate browsing, searching, sorting, and scanning of articles. Information extraction is also used in novel ways to provide high level views of multiple documents, for example, by presenting the most frequently mentioned diseases each day. MiTAP currently supports users in the medical and humanitarian relief areas. We attribute MiTAP's success to the careful crafting of useful features from imperfect human language component technologies. We will discuss how these natural language technologies have been used to create and enhance the prototype system, the shortcomings and limitations of those technologies, and planned extensions and improvements to help users deal with overwhelming amounts of data and information in the biomedical domain.

Project Argus

James Wilson, V, MD, (ISIS Center, Georgetown University) IceAxe5@aol.com

Abstract

The frequency of biological events relevant to national security is increasing, and current US disease surveillance systems have demonstrated failure to detect these events in a timely fashion. Examples include the HIV/AIDS pandemic, the escape of Rift Valley Fever from Africa, the foot and mouth disease pandemic, West Nile virus translocation to the US, post-9/11 anthrax attacks, SARS, and recently, the translocation of monkeypox to the US. The clear and present danger to the US spans human, animal, and plant infectious disease.

Current biosurveillance by the United States is heavily focused on peri-event markers for epidemic activity gleaned almost exclusively from human medical surveillance data. These peri-event markers (e.g., positive biosensor detection per BioWatch) typically provide responders lead-time to response on the order of days. Surveillance systems are ,stovepiped, across multiple agencies with no integrative alerting capability, contributing to the lack of rapid, timely alerting. Much of the current expenditure in biodefense research is focused on response capability without balanced consideration of enhancing the current national surveillance capability.

Indications and Warnings (I&Ws) provide the capability to alert US responders of an imminent biological event weeks to months in advance and thus are considered pre-event markers. In effect, I&Ws have the potential to ,prime, the national response infrastructure by alerting agencies of an evolving threat that could ultimately be catastrophic. Sources of I&Ws include human, veterinary, zoological park, and insect vector surveillance; media; internet and telecommunication traffic; commerce; economic indices; ,farm to fork, agricultural and food production surveillance; and ground climatological and remotely sensed environmental indices. Retrospective analyses of major biological events such as West Nile Virus and SARS have revealed multiple I&Ws were present weeks to months in advance in multiple disparate data sources but were unable to be recognized and utilized properly by the national response community.

Key obstacles to I&W research, development, and integration include lack of recognition by the homeland security and public health communities due to the disparate data sources needed for the analysis. Monitoring I&Ws requires robust situational awareness by the analyst and broad capability to cross multiple traditional mission statements and scientific disciplines. Further, data security and assurance is complex when considering the number of data streams involved that are unclassified, proprietary, and classified. Once I&Ws enter the classified environment, it is difficult to mobilize the information in a timely fashion to local responders. However, solutions are currently being investigated.

Integration of I&Ws in the biosurveillance strategy is a critical need for US biosecurity. Given the increased frequency of bioevents over the past 20 years, an integrative strategy that spans the human and agricultural biodefense concerns is a necessity.

Thursday 1:45 p.m. to 3:30 – Computational Geometry and Robust Statistics (Invited Session)

On the Least Median Square Problem

David M. Mount, (University of Maryland), mount@cs.umd.edu,
Jeff Erickson, (University of Illinois), jeffe@cs.uiuc.edu, and
Sariel Har-Peled, (University of Illinois), sariel@cs.uiuc.edu

Abstract

We consider the exact and approximate computational complexity of the multivariate LMS linear regression estimator. The LMS estimator is among the most widely used robust linear statistical estimators. Given a set of n points in d-dimensional space and a parameter k, the problem is equivalent to computing the slab bounded by two parallel hyperplanes of minimum separation that contains k of the points. We present algorithms for the exact and approximate versions of the multivariate LMS problem. We also provide nearly matching lower bounds on the computational complexity of these problems. The lower bounds hold if deciding whether d+1 points are coplanar requires $\Omega(n^d)$ time.

Computational Geometry and Statistical Depth measures

Eynat Rafalin, (Tufts University), erafalin@cs.tufts.edu, and
Diane L. Souvaine, (Tufts University), dls@cs.tufts.edu

Abstract

The computational geometry community has long recognized that there are many important and challenging problems that lie at the interface of geometry and statistics. The relatively new notion of data depth for non-parametric multivariate data analysis, is inherently geometric in nature, and therefore provides a fertile ground for expanded collaboration between the two communities.

Over a decade ago, point-line duality and combinatorial and computational results on arrangements of lines contributed to the development of an efficient algorithm for two-dimensional computation of the LMS regression line. The same principles and refinements of them are being used today for more efficient computation of data depth measures such as halfspace (Tukey) depth or regression depth. Other depth measures, such as simplicial depth, pose open research problems in the computational geometry community. Furthermore, other problems are solved in theory by algorithms with efficient asymptotic running times, where, in practice, the hidden constants make implementations infeasible, prompting the need for continued research and implementable solutions.

New developments and increased emphasis in the area of multivariate analysis heighten the need for new and efficient computational tools and for an enhanced partnership between statisticians and computational geometers. Much is to be gained by increased collaboration and solutions that are not only provably efficient but also effective in practice.

Imputation of Microarray Data

Ming Ouyang, (University of Medicine and Dentistry of New Jersey), ouyangmi@umdnj.edu

Abstract

In microarray experiments, missing entries arise from blemishes on the chips. In large-scale studies, virtually every chip contains some missing entries and more than 90

Two evaluation metrics of imputation accuracy are employed. First, the root mean squared error measures the difference between the true values and the imputed values. Second, the number of mis-clustered genes measures the difference between clustering with true values and that with imputed values; it examines the bias introduced by imputation to clustering. The Gaussian mixture clustering with model averaging imputation is superior to all other imputation methods, according to both evaluation metrics, on both time-series (correlated) and non-time series (uncorrelated) data sets.

Thursday, 1:45 p.m. to 3:30 – Microarrays II (Contributed)

Rank-Based Classification of Gene Expression Profiles

Donald Geman, (Johns Hopkins University), geman@jhu.edu,
Daniel Naiman, (Johns Hopkins University), daniel.naiman@jhu.edu, and
Christian d'Avignon, (Johns Hopkins University), davic@ bme.jhu.edu

Abstract

Statistical inference from gene expression microarray data is difficult due to the small number of observations, typically tens, relative to the large number of genes, typically thousands. Consequently, standard methods in machine learning may lead to over-fitting and inflated estimates of performance in detecting disease, identifying tumors and predicting treatment responses, especially when all aspects of learning a classifier are not properly cross-validated. Moreover, and equally important, the results may be very difficult to interpret in biological terms. We address these problems by a purely rank-based analysis, for instance comparing the mRNA counts in selected pairs. As an example, we attempt to distinguish among cancer types with a maximum likelihood classifier based on a single pair of genes. The results so far are very promising; we obtain accurate and transparent decisions from small samples in standard classification tasks. However, there are many unanswered questions, both statistical and biological.

A Comparison of Direct and Sequential False Discovery Rate Algorithms: Computational Experiments for Exploratory DNA Microarray Studies

Danh V. Nguyen, (University of California, Davis), ucdnguyen@ucdavis.edu

Abstract

The problem of detecting differential gene expression with microarray data has led to innovative approaches to controlling false positives in multiple testing. False discovery rate (FDR) has been widely used as a measure of error in this multiple testing context. Direct estimation of FDR was recently proposed by Storey (2002) as a substantially more powerful alternative to the traditional sequential FDR controlling procedure, pioneered by Benjamini and Hochberg (1995). Direct estimation to FDR requires fixing a rejection region of interest and then conservatively estimating the associated FDR. On the other hand, sequential FDR procedure requires fixing a FDR control level and then estimating the rejection region. Thus, sequential and direct approaches to FDR control appear fundamentally different. However, these approaches can be unified and the methods compared using computational experiments designed to more reflect exploratory DNA microarray studies often implemented in practice. Using simulation, we illustrate that modified sequential FDR algorithms are equivalent to the direct estimates of FDR and, hence, are as powerful. In addition, both approaches simply approximate the least conservative (optimal) sequential FDR procedure.

Probabilities of Spurious Connections in Gene Networks: Application to Expression Time-Series

David R. Bickel, (Medical College of Georgia), bickel@prueba.info

Abstract

Motivation: The reconstruction of gene networks from gene expression microarrays is gaining popularity as methods improve and as more data become available. The reliability of such networks could be judged by the probability that a connection between genes is spurious, resulting from chance fluctuations rather than from a true biological relationship. Results: Unlike the false discovery rate and positive false discovery rate, the decisive false discovery rate (dFDR) is equal to a conditional probability without assuming weak dependence or the randomness of hypothesis truth values. This property is useful not only in the common application to the detection of differential gene expression, but also to the problem of determining the probability of a spurious connection in a reconstructed gene network. Estimators of the dFDR can estimate each of three probabilities: 1. The probability that two genes that appear to be associated with each other lack such association. 2. The probability that a time ordering observed for two associated genes is misleading. 3. The probability that a time ordering observed for two genes is misleading, either because they are not associated or because they are associated without a lag in time. The first probability applies to both static and dynamic gene networks, and the other two only apply to dynamic gene networks. Availability: Cross-platform software is free from http://www.davidbickel.com as R source code and a Java application.

DNA Microbial and Viral Identification using Ultraspecific Probes "Blind" to Host and Background DNA

Catherine Putonti, (University of Houston), putonti@bioinfo.uh.edu,
George E. Fox, (University of Houston), fox@uh.edu,
Richard C. Willson, (University of Houston), willson@uh.edu, and
Yuriy Fofanov, (University of Houston), yfofanov@bioinfo.uh.edu

Abstract

The reliable detection and identification of microbes and viruses in complex samples without separation of DNA of the organism of interest from the sample background is a challenging and important problem. We have developed a set of novel algorithms that make it feasible to analyze the occurrence of all possible short sequences of length 10 to 25 nucleotides in complete genome sequences of any size. As a result, we can identify all signature sequences present in each of a large set of pathogen genomes and absent in (and not within up to three mismatches) the human genome. We found that it is unusual to find a single, unique genomic sequence present simultaneously in all genomes of interest and absent in all other genomes, including the host organism, even for groups of closely related organisms (e.g., the West Nile virus). This result leads us to suggest using a set of probes that are absent in the host genome, likely to be found in the pathogen genome, and expressed in a unique pattern for each pathogen for pathogen identification. Herein we use an evolutionary programming approach to design microarrays so as to minimize the number of probes required, to avoid false positives and to achieve maximal sensitivity. Supporting the proposed approach, initial in silico and in vitro microarray experimental results are provided.

Thursday 4:00 p.m. to 5:45 – Functional Data Analysis for Computational Biology (Invited Session)

Self Modeling with Flexible, Random Time Transformations

Lyndia C. Brumback, (University of Washington), lynb@u.washington.edu, and
Mary J. Lindstrom, (University of Wisconsin), lindstro@biostat.wisc.edu

Abstract

Methods for modeling sets of complex curves where the curves must be aligned in time (or in another continuous predictor) fall into the general class of functional data analysis and include self modeling regression and time warping procedures. Self modeling regression (SEMOR), also known as a shape invariant model (SIM), assumes the curves have a common shape, modeled non-parametrically, and curve-specific differences in amplitude and timing, traditionally modeled by linear transformations. When curves

contain multiple features that need to be aligned in time, SEMOR may be inadequate since a linear time transformation generally cannot align more than one feature. Time warping procedures focus on timing variability and on finding flexible time warps to align multiple data features. We draw on these methods to develop a SIM that models the time transformations as random, flexible, monotone functions.

Modeling Continuous Shape Change for Facial Animation

Julian Faraway, (University of Michigan), faraway@umich.edu

Abstract

The movement of landmarks on the human face can be recorded in 3D using motion capture equipment. We describe methods for the analysis of data collected on groups of subjects with a view to describing and assessing the differences between the facial motions of those groups. We focus on the smile motion in particular. The methods presented can be used more generally for continuous shape change data.

We introduce a novel parameterization of shape change that allows the parsimonious description of facial motion. We allow for a distinction between static facial shape and dynamic facial motion. We describe statistical methods for modeling differences in facial motion including a comparison of mean motions, principal components for describing the variation in motion and linear models for describing the effects of predictors.

Detecting Changes in Brain Shape, Scale and Connectivity via the Geometry of Random Fields

Keith Worsley, (McGill University), keith.worsley@mcgill.ca

Abstract

Three types of data are now available to test for changes in brain shape: 3D binary masks, 2D triangulated surfaces, and trivariate 3D vector displacement data from the non-linear deformations required to align the structure with an atlas standard. We use the Euler characteristic of the excursion set of a random field as a tool to test for localized shape changes. We extend these ideas to scale space, where the scale of the smoothing kernel is added as an extra dimension to the random field. Extending this further still, we look at fields of correlations between all pairs of voxels, which can be used to assess brain connectivity. Shape data is highly non-isotropic, that is, the effective smoothness is not constant across the image, so the usual random field theory does not apply. We propose a solution that warps the data to isotropy using local multidimensional scaling. We then show that the subsequent corrections to the random field theory can be done without actually doing the warping - a result

guaranteed in part by the famous Nash Embedding Theorem. This has recently been formalized by Jonathan Taylor who has extended Robert Adler's random field theory to arbitrary manifolds.

Thursday, 4:00 p.m. to 5:45 – Future of Statistical Software (Invited Session)

Challenges for Future Statistical Software for Non-expert Users

John Sall, (SAS Institute), sall@sas.com

Abstract

Statistical software has evolved a lot, but it still has a long way to go before it is truly friendly to the non-expert user. We have lots of computing power, and lots of tools, but they are not easy to use. We have poor generalizability of standard statistical methods. Our tools often don't scale well to large problems. We have weaknesses in operating system support for very basic needs such as scalable graphics formats, and the ladder of help. Document architecture is not rich enough. Our teaching and our cookbooks don't cover enough.

Yxilon – Designing the Next Generation, Vertically Integrable Statistical Computing Environment

Uwe Ziegenhagen, (HU Berlin), ziegenhagen@wiwi.hu-berlin.de, and
Prof. Dr. W. Haerdle, (HU Berlin), haerdle@wiwi.hu-berlin.de

Abstract

Modern statistical computing requires smooth integration of new algorithms and quantitative analysis results into all sorts of platforms, like webbrowsers, standard and proprietary applications.

With Yxilon we want to implement such a vertically integrable, modular environment, providing the user with a rich set of statistical methods and a variety of different interfaces to use these methods.

Yxilon will be the successor of XploRe, a complete statistical engine developed at Humboldt-Universitaet zu Berlin. While working on several projects with international partners the function set of XploRe had been more and more extended at the cost of performance and stability, aspects we want to change in the upcoming Yxilon.

The main goals of Yxilon are:

- platform independence
- COM and Client/server interfaces
- database functionality and connectivity
- multi-lingual user interfaces
- full support of all XploRe functions and packages
- Support for XML

XML-Based Applications in Statistical Analysis

Yuichi Mori, (Okayama University of Science), mori@soci.ous.ac.jp,
Tomokazu Fujino, (Fukuoka Women's University), fujino@fwu.ac.jp,
Yoshiro Yamamoto, (Tama University), yama@tama.ac.jp, and
Tomoyuki Tarumi, (Okayama University of Science), t2@ems.okayama-u.ac.jp

Abstract

In the Internet age, many things have become more web-intensive than before and statistics is no exception to this trend. On the other hand, it has been desirable that statistical information is provided in some unified format. Concerning these situations, XML takes an important and useful role since XML is designed for large-scale electronic publishing and the exchange of a wide variety of data on the Web and elsewhere and also many tools using XML technologies are currently provided.

In this paper the following XML-based applications that we are now developing using XML are introduced: Database of datasets and analysis stories which are written in XML; On-line analysis system in which automatic analyses can be performed with initial parameters provided from XML documents; Web-based statistical graphics and map using SVG (Scalable Vector Graphics) and X3D (eXtensible 3D); and On-line interactive text using XML technologies. These applications will be considered from the aspects such as data-description, accessibility to statistical engines, documentation and interactivity.

Thursday, 4:00 p.m. to 5:45 – The Best of Data Mining from KDD (Invited Session)

Mining Concept-Drifting Data Streams Using Ensemble Classifiers

Haixun Wang, (IBM T. J. Watson Research Center), haixun@us.ibm.com

Abstract

Recently, mining data streams with concept drifts for actionable insights has become an important and challenging task for a wide range of applications including credit card fraud protection, target marketing, network intrusion detection, etc. Conventional knowledge discovery tools are facing two challenges, the overwhelming volume of the streaming data, and the concept drifts. In this paper, we propose a general framework for mining concept-drifting data streams using weighted ensemble classifiers. We train an ensemble of classification models, such as C4.5, RIPPER, naive Bayesian, etc., from sequential chunks of the data stream. The classifiers in the ensemble are judiciously weighted based on their expected classification accuracy on the test data under the time-evolving environment. Thus, the ensemble approach improves both the efficiency in learning the model and the accuracy in performing classification. Our empirical study shows that the proposed methods have substantial advantage over single-classifier approaches in prediction accuracy, and the ensemble framework is effective for a variety of classification models.

Mining Distance-Based Outliers in Near Linear Time

Stephen Bay, (Stanford University), sbay@apres.stanford.edu, and
Mark Schwabacher, (NASA Ames Research Center), mark.a.schwabacher@nasa.gov

Abstract

Defining outliers by their distance to neighboring examples is a popular approach to finding unusual examples in a data set. Recently, much work has been conducted with the goal of finding fast algorithms for this task. We show that a simple nested loop algorithm that in the worst case is quadratic can give near linear time performance when the data is in random order and a simple pruning rule is used. We test our algorithm on real high-dimensional data sets with millions of examples and show that the near linear scaling holds over several orders of magnitude. Our average case analysis suggests that much of the efficiency is because the time to process non-outliers, which are the majority of examples, does not depend on the size of the data set.

Privacy Preserving K-Means Clustering over Vertically Partitioned Data

Jaideep Vaidya, (Purdue University), jsvaidya@cs.purdue.edu, and
Chris Clifton, (Purdue University), clifton@cs.purdue.edu

Abstract

Privacy and security concerns can prevent sharing of data, derailing data mining projects. Distributed knowledge discovery, if done correctly, can alleviate this problem. The key is to obtain valid results, while providing guarantees on the (non)disclosure of data. We present a method for k-means clustering when different sites contain different attributes for a common set of entities. Each site learns the cluster of each entity, but learns nothing about the attributes at other sites.

Thursday, 4:00 p.m. to 5:45 – Computational Statistics I (Contributed Session)

Estimating the Parameters of Infinite Scale Mixtures of Normals

Hasan Hamdan, (James Madison University), hamdanhx@jmu.edu, and
John Nolan, (American University), Jpnolan@american.edu

Abstract

Conditions and classes of examples of variance mixture of normals are given, along with a constructive proof on how to guarantee that a finite variance mixtures of normals is uniformly close (up to a desired tolerance level) to a given infinite variance mixture distribution.

We wish to minimize the finite number of terms needed subject to a specified desired tolerance level. The method, which is based on discretizing the mixing measure is presented and illustrated through an example and the infinite and finite mixtures are displayed on the same graph. A new method for estimating the parameters of a variance mixture of normals is also introduced. The new method is based on minimizing the squared distance between the estimated density and the corresponding density computed by discerizing the mixture over a predetermined grid of R values and a grid of X values. This method looks promising especially for modeling data.

Permutation Tests in Assessing Survival Forests for Prognosis Based on Gene Profiles

Van L. Parsons, (National Center for Health Statistics), vparsons@cdc.gov, and
Thu M. Hoang, (Universite Rene Descartes), hoang@biomedicale.univ-paris5.fr

Abstract

Combinations of survival regression trees called survival forests (SF) applied to microarray data provide both prediction of individual survival functions and the corresponding ranking of variable importance although without assessment for the latter. A basic

question is whether the structures of SF and resulting statistics can be attributed to chance alone. For small data sets we propose the use of permutation tests as a way to determine significance of the SF performance, both for assessment of the fit and to test the significance of the genes identified as prognostic markers.

Wavelet Domain Linear Inversion via the LASSO

Leming Qu, (Boise State University), qu@math.boisestate.edu, and
Partha Routh, (Boise State University), routh@cgiss.boisestate.edu

Abstract

We propose an approach for solving the linear inverse problems in the wavelet domain. The solution minimizes the residual sum of squares subject to the L_1 norm of the Discret Wavelet Transform of the underlying signal less than a constant. Such solution in wavelet domain is sparse and is more appropriate for nonsmooth signals. We use the "Least Absolute Shrinkage and Selection Operator" (LASSO) modification of the "Least Angle Regression" (LARS) algorithm proposed by Efron et. al. (2004) to carry out the numerical computation. A simulation result confirmed the good performance of the approach for the optimally selected tuning parameter.

Cramér-Rao Bounds and Monte Carlo Calculation of the Fisher Information Matrix in Difficult Problems

James C. Spall, (Johns Hopkins Univ., APL), james.spall@jhuapl.edu

Abstract

The Fisher information matrix summarizes the amount of information in the data relative to the quantities of interest. There are many applications of the information matrix in modeling, systems analysis, and estimation, including confidence region calculation, input design, prediction bounds, and noninformative priors for Bayesian analysis. This paper reviews some basic principles associated with the information matrix, presents a resampling-based method for computing the information matrix together with some new theory related to efficient implementation, and presents some numerical results. The resampling-based method relies on an efficient technique for estimating the Hessian matrix, introduced as part of the adaptive (second-order) form of the simultaneous perturbation stochastic approximation (SPSA) optimization algorithm.

Key words: Monte Carlo simulation; Cramr-Rao bound; simultaneous perturbation; antithetic random numbers.

User Profiling in Window Title and Process Table

Chien-Chih Lin, (School of Computational Sciences, George Mason University), clin3@gmu.edu,
Eun Young Noh, (School of Computational Sciences, George Mason University), enoh@gmu.edu,
Youngping Yan, (School of Information Technology and Engineering, George Mason University), yyan1@gmu.edu, and
Dr. Edward Wegman, (School of Information Technology and Engineering, George Mason University), ewegman@galaxy.gmu.edu

Abstract

As one of strategies of computer security, user profiling was conducted for the window title and process table data which was collected on an internet connected unclassified window NT network reflecting the currently ubiquitous window based computing environment. The statistical methodology in the NIDES(Next Generation Detection Expert System) was implemented on these data to profile the past behavior of users on a computer system and to learn the pattern of each user and to identify the unauthorized user or misuse of authorized user. The anomaly detection was based on the score value that represents how much the users behavior is abnormal compared to the past patterns of behavior. As a method of measuring performance, a specified user was assumed as an authorized user and the other users were treated as unauthorized users. The score values of a specified user and other users gave quite different distributions and the rate of detecting other users from a specified user was about 60 70

Thursday, 4:00 p.m. to 5:45 – Clustering (Contributed Session)

Cluster Subspace Identification Via Conditional Entropy Calculations

James Diggans, (George Mason University), jdiggans@gmu.edu, and
Jeffrey L. Solka, (George Mason University), jsolka@gmu.edu

Abstract

Methods of high-level data exploration capable of robustness in the face of noise found within microarray data are few and far between. Solutions making use of all original features to derive cluster structure can be misleading while those that rely on a trivial feature selection can miss important characteristics. We present a method adapted from previous work in the field of geography (Guo et al, Wrokshop on Clustering High Dimensional Data and its Applications 2003) relying upon conditional entropy between pairs of dimensions to uncover underlying, native cluster structure within a dataset. Results will be presented on artificial and gene expression data sets.

Actor Allegiance and Blockmodel Strength

John Rigsby, (Naval Surface Warfare Center), RigsbyJT@nswc.navy.mil, and
Jeffrey L. Solka, (Naval Surface Warfare Center), SolkaJL@nswc.navy.mil

Abstract

This presentation examines a new method for testing blockmodel strength. The new method is called allegiance. Allegiance is a measure of how much an actor is helping his block.

Our intent is to measure blockmodel strength change by examining actor allegiance as the partition size changes. Ultimately we would like to use our developed approach as a way to automatically calculate the partition size of the proper blockmodel.

Bayesian Hierarchical Models of the Browsing Behavior of World Wide Web Users

Juana Sanchez, (University of California Los Angeles), jsanchez@stat.ucla.edu, and
Ching-Ti Liu, (University of California Los Angeles), ctliu@stat.ucla.edu

Abstract

We consider the case of surfing within a single large Web site, which is important from the point of view of site design, web server proxy efficiency and search engine optimal ranking of pages. The site used as an example to illustrate the methods is msnbc.com. We use a set of server log data on the Web pages chosen by 989818 users in a twenty-five hour period, where the response measure for each user is an ordered sequence of choices among 17 categories (UCI KDD Archive). A common way to model the browsing behavior of users is to assume that the decision of users is a random walk with a probability distribution of first passage time to a threshold that is a two-parameter inverse-gaussian distribution. Another hypothesis examined is that users at each page conduct an independent Bernoulli trial to make a stopping decision, which implies a geometric distribution. Mixtures of first-order markov processes or model-based clustering with and without a Bayesian flavor have offered very useful exploratory data analyses. All these studies have shown evidence that web-surfing behavior may be non-Markov in nature and have illustrated how hard it is to capture dependencies in the data; the performance of the models over a wide range of Web Site formats is still inconclusive. This performance has been measured by the ability to predict page hits, by the resulting distribution of page hits, and by the contribution to efficient web caching schemes. Some models have been tested with server log data of AOL or similar Sites and others have been tested within a single Web site like msnbc.com. The levels of aggregation of pages and clustering of user behavior have also varied within studies. In this paper, we argue that for the case of browsing within a news portal like msnbc.com, where contents are continually changing, the

server-log data is only meaningful when categories are aggregated, like they are for the msnbc.com data set, and the order of the browsing may not be relevant. We use a complex Bayesian hierarchical model of the page counts per user. This model has the ability to have enough parameters to fit the data well, while using a population distribution to structure dependence in the parameters. The model can be generalized to different types of Web sites, different levels of aggregation of pages and different clustering schemes. We compare the performance of this new model to that of previous models.

Model-based Clustering With an Adaptive Mixtures Smart Start

Jeffrey L. Solka, (NSWCDD), solkajl@nswc.navy.mil, and
Wendy L. Martinez, (ONR), Wendy_Martinez@onr.navy.mil

Abstract

This talk will discuss a new model-based clustering methodology. This approach is predicated on the use of a recursive semi-parametric density estimation procedure, the adaptive mixture method of Priebe, as the starting point to the agglomerative phase of the model-based clustering procedure. The computational efficiency of the model-based clustering methodology is improved in that one does not have to perform the agglomerative phase of the procedure using the complete data set. The methodology along with recent results obtained on artificially generated data will be presented.

Friday, 8:00 a.m. to 9:45 a.m. – Best of the IASC I (Invited Session)

Using Bayesian Networks to Reconstruct Yeast Genetic Networks

Grace S. Shieh, (Institute of Statistical Science, Academia Sinica, Taiwan),
gshieh@stat.sinica.edu.tw

Abstract

A dynamic Bayesian networks approach is developed to reconstruct DNA synthesis and repair genetic networks in yeast. This approach using Lamakian genetic algoritm to optimize the structure search and simulated annealing to estimate the parameters involved.

We simulated data from a non-linear function with parameters trained from real mciroarray data (Spellman et al., 1998). This simulated data were sued to test and train the performances of the Bayesian networks approach.Finally, we applied the algorithm to real data sets in Spellman et al. (1998).

Does Sequence Similarity Predict Expression Similarity

Kui Zhang, (Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham), kzhang@ms.soph.uab.edu,
David Allison, (Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham), DAllison@ms.soph.uab.edu,
Grier Page, (Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham), gpage@ms.soph.uab.edu, and
Elliot J. Lefkowitz, (Department of Microbiology, University of Alabama at Birmingham), elliotl@uab.edu

Abstract

With the completion of the human genome project and the rapid development of high throughput technologies, such as microarrays, yeast-two-hybrid systems, large-scale deletion experiments, etc., a large amount of biological data are now available and certainly more data will be generated in the near future. Information extracted from combined data can provide important insight into the function of genes and better understanding of the genetic networks underlying complex biological processes. For example, microarray data has been used combining with other kinds of data, such protein-protein interaction data to improve the prediction of the function of genes. In this paper, we ask if sequence similarity of genes predicts similarity in expression pattern. We identify genes with similar sequence by BLAST searches and determine the differentially expressed genes from microarray experiments. We develop a method to combine these two related sets of information.

Statistical Methods for Proteomics

Francoise Seillier-Moiseiwitsch, (University of Maryland Baltimore County), seillier@math.

Abstract

The term 'proteome' has been coined to reflect the revolutionary changes the field of biochemistry has been undergoing. This word refers to the PROTEins expressed by a genOME or tissue. Unlike the genome, the proteome is affected by a number of factors such as tissue and environmental conditions. A gene can be spliced in many different ways and proteins can be altered after translation. Hence, the proteome consists of far more proteins than the genome contains genes. Two-dimensional polyacrylamide gel electrophoresis is currently the technique of choice to separate and display all the proteins expressed in a tissue. In the resulting protein map, individual protein spots are identified and quantified. To this end, background noise and dust spikes are eliminated. Wavelets are utilized to summarize the two-dimensional gels. Within this framework, we seek to compare different gels by determining the strength and the location of the signal.

Friday, 8:00 a.m. to 9:45 a.m. – Mass Spectroscopy and Clinical Proteomics (Invited Session)

Signal Conditioning and Filtering of SELDI Mass Spectrometry Time Series

Dariya Malyarenko, (Applied Science, College of William and Mary and INCOGEN, Inc.), dasha@compsci.wm.edu,
Dennis Manos, (Applied Science, College of William and Mary), dmanos@as.wm.edu,
William Cooke, (Physics, College of William and Mary), wecook@wm.edu, and
Eugene Tracy, (Physics, College of William and Mary), ertrac@wm.edu

Abstract

Current clinical and biological research is devoted to mass spectrometric measurement of peptide changes in body fluids that may indicate the presence of disease. Time-of-flight (TOF) mass spectrometry, like SELDI, digitally records the voltage output from ion detectors connected to electronics designed to provide low noise, wide dynamic range, high sensitivity, and repeatable response. Understanding the physical limitations of such electronics is necessary to develop algorithms required to suitably adjust data to account for instrumental effects in the acquired raw records prior to subsequent statistical data analysis and interpretation. We address these physical limitations and discuss their resulting unwanted attributes in SELDI TOF data records, including time dependent gain, baseline shifts, peak overlaps, and time-domain jitter. We describe methods to detect and correct such instrumental effects, and computational algorithms to enhance the mass resolution in corrected raw data, using time series analysis and filtering techniques.

A Multiresolution View of Protein Mass Spectrometry Data

Timothy W. Randolph, (University of Washington), trandolp@fhcrc.org, and
Yutaka Yasui, (Fred Hutchinson Cancer Research Center), yyasui@fhcrc.org

Abstract

In current research on finding biomarkers for disease, the use of high-throughput mass spectrometry data has become common. These data consist of a very wide range of peptide/protein masses under a few thousand Daltons. The analysis of such data involves inferring the existence of a peptide/protein of a particular mass from the existence of a spike (or peak) in the spectrum output of a mass spectrometer. In using these spectra for identifying biomarkers, many data-analytic questions arise: What constitutes the noise within a spectrum? What defines a peak? Since a peak mass for an identical peptide/protein appears at slightly shifted mass points across samples, how should one treat peaks at close mass points? What is the true mass of a candidate-marker peptide/protein? In view of large variations in peak heights, what

role should the intensity measure of the spectrum in the analysis? In particular, should the intensity measurements across samples be normalized? Should one adjust for the trend in a spectrum; i.e., should a baseline curve be subtracted from the spectrum before considering any of these questions, and if so, how?

Our work uses wavelet methods to form a multiresolution view of the spectrum data and focus on local structure occurring at various scales. It provides an objective method for processing these data that avoids ad hoc decisions in addressing the above questions. In particular, the method involves no normalization procedure or baseline removal, and does not define peaks based on a direct calculation of signal-to-noise ratios. The application of our method will be illustrated using data from a biomarker discovery study in early cancer detection research .

Friday, 8:00 a.m. to 9:45 a.m. – Text Mining and Applications (Invited Session)

Identifying Cross Copora Document Associations Via Minimal Spanning Trees

Jeffrey L. Solka, (NSWCDD), solkajl@nswc.navy.mil,
Avory Bryant, (NSWCDD), bryantac@nswc.navy.mil, and
Edward J. Wegman, (George Mason University), ewegman@galaxy.gmu.edu

Abstract

This talk will focus on our recent work in the identification of related documents from different corpora or discipline areas. The purpose of this work is to help a researcher or program manager to identify fruitful cross-disciplinary research areas. The talk will discuss some preliminary results that have been obtained using a small Science News (around 1200 articles) dataset. The work is predicated on new visualization environment to facilitate the exploration of the multi-class interpoint distance matrix. This work is joint with the Algotek Team in general and Edward J. Wegman of GMU and Avory Bryant of NSWCDD specifically.

Intersection Graphs for Text Analysis

Elizabeth Leeds, (NSWC), leedsem@nswc.navy.mil, and
David J. Marchette, (NSWC), marchettedj@nswc.navy.mil

Abstract

In text analysis, it is often the case that each document is represented by a vector in high dimensional space where the dimension is equal to the number of words in

the lexicon. The features extracted for document processing are weights on the words, where the weights are corpus dependent. In this work, we use random graphs to analyze the results of feature selection in text processing applications. We define the vertex set as the set of all documents and the edge set is defined in terms of the number of words in common between documents. We discuss methods for utilizing the intersection graphs to study the effect of dimensionality reduction, corpus changes, and document changes on the classification rate.

Document Classification and Clustering Using Weighted Text Proximity Matrices

Wendy L. Martinez, (Office of Naval Research), martinwe@onr.navy.mil,
Angel R. Martinez, (NSWCDD), martinezar@nswc.navy.mil, and
Edward J. Wegman, (George Mason University), ewegman@gmu.edu

Abstract

In previous work, we introduced a way of encoding free-form text documents called the bigram proximity matrix (BPM). When this encoding was used on a corpus of documents, where each document is tagged with a topic label, results showed that the documents could be classified based on their tagged meaning. In this paper, we develop and investigate various methods of weighting the elements of the BPM, which are analogous to the weighting schemes found in natural language processing and information retrieval. These include logarithmic weights, augmented normalized frequency, inverse document frequency and pointwise mutual information. We apply k-nearest neighbor classification and model-based clustering to assess the performance of the weighted BPMs. Results presented in this paper show that some of the weights improved the classification performance.

Friday, 8:00 a.m. to 9:45 a.m. – Genetic Algorithms for Computational Biology (Invited Session)

Visual Data Mining of RNA Secondary Structure Folding Pathways as Determined by the Massively Parallel Genetic Algorithm

Bruce A. Shapiro, (Laboratory of Experimental and Computational Biology, NCI-Frederick), bshapiro@ncifcrf.gov, and
Wojceich Kasprzak, (Basic Research Program, SAIC-Frederick), Kasprzak@ncifcrf.gov

Abstract

RNA folding pathways are proving to be quite important in the determination of RNA function. Studies indicate that RNA may enter intermediate and multiple conformational states that are key to its functionality. These states may have a significant impact on gene expression and molecular function. It is known that the biologically functional states of RNA molecules may not correspond to their minimum energy state, that kinetic barriers may exist that trap the molecule in a local minimum, that folding often occurs during transcription, and that cases exist in which a molecule will transition between one or more functional conformations. Thus, methods for simulating the folding pathway and dynamic behavior of an RNA molecule are important for the prediction of RNA structure and its associated function.

We have developed several visual data mining techniques associated with a massively parallel genetic algorithm for RNA structure prediction, as well as with STRUCTURELAB, our RNA structure analysis workbench. These methodologies are used to determine the significant intermediate and final structures associated with RNA folding. Since the genetic algorithm is essentially stochastic, multiple runs are required. The visualization procedures used give significant feedback concerning the characteristics of the folding runs. This feedback includes: interpretation of results from individual genetic algorithm runs that are based on population consensus or best fit structures, this includes the discovery of transition states in the folding process; final results of individual runs; and the interpretation of genetic algorithm results from multiple RNA sequences from the same family to identify common structural elements across the different sequences. In addition, fitness maps as well as results derived from different population sizes are used.

The combination of the visualization techniques as well as other methodologies embedded within the STRUCTURELAB and genetic algorithm environments help to determine the overall picture representing the folding pathway or final structure(s) of a given RNA sequence. This paper will describe several of these techniques and show how they are used to help solve this very highly combinatoric problem.

Knowledge Discovery in Large Biological Data Sets Using Hybrid Classifier/Evolutionary Algorithms

Michael Raymer, (Wright State University), mraymer@cs.wright.edu,
Mike Peterson, (Wright State University), peterson.7@wright.edu, and
Travis Doom, (Wright State Unversity), doom@cs.wright.edu

Abstract

A key element of bioinformatics research is the extraction of meaningful information from large experimental data sets. A variety of pattern recognition and machine learning methods have been applied to this task, some exhibiting significant success for specific data sets. We have, in the past, shown that an evolutionary algorithm (EA) can be used to extract features from large biological data sets. These EA-based feature extractors have performed well with a variety of classifiers, including Euclidean

distance-based k-nearest-neighbor classifiers and the Naive Bayes classifier. Recently, we have explored the use of the cosine-similarity distance metric in combination with the k-nearest-neighbors classifier and EA feature extraction. Using the evolutionary algorithm to evolve feature weights and coordinate system displacement, we have found the EA-cosine-knn hybrid classifier to be an effective tool for classification of a variety of biological data sets including protein solvation and identification of gene response signatures from RNA expression data.

Polyoptimizing Genetic Algorithms for Feature Selection

Ewy Mathe, (George Mason University), emathe@gmu.edu, and
John Grefenstette, (George Mason University), jgrefens@gmu.edu

Abstract

The analysis of large biological data sets that arise in gene expression or proteomics experiments often involves the selection of a subset of the available features that supports efficient classification. Finding multiple, distinct solutions to the feature subset selection problem may lead to increased biological insights. In this talk we address the problem of finding multiple solutions to the feature subset selection problem using a polyoptimizing genetic algorithm which incorporates a dynamic penalty function. We illustrate the approach on an ovarian cancer classification problem using proteomics data.

Friday, 8:00 a.m. to 9:45 a.m. – Classification (Contributed Session)

Comparison of Classification Techniques in Bioinformatics

Rashpal Ahluwalia, (West Virginia University), rashpal.ahluwalia@mail.wvu.edu, and
Sundar Chidambaram, (West Virginia University), schidam@mix.wvu.edu

Abstract

This paper provides a comparison of Discriminant Function Analysis (DFA), Logistic Regression, Decision Tree, and Artificial Neural Network algorithms utilized in bioinformatics. DFA is used to predict group membership in naturally occurring groups. Its goal is to find a dimension along which the groups differ. It is used when the Dependent Variable (DV) is predicted from a set of Independent Variables (IVs). The prediction success of DFA is determined by the choice of the predictors. DFA assumes the predictors to be normally distributed and linearly related. DFA provides accurate predictions when the group sizes are equal and when IVs are continuous and well distributed.

Logistic Regression (LR) allows the prediction of a DV from a set of IVs that may be discrete, continuous, or a mix. The models produced by LR are non-linear. Unlike DFA, LR makes no assumptions on the predictor variables. It predicts the probability of a particular outcome for each sample. It is also robust for complex datasets. Decision Trees (DT) are generally used to predict discrete valued outputs. DTs can generally be represented by a set of if-then rules. DTs are suitable when the instances are represented by disjoint values and when the training data contain errors or has missing values.

The classification algorithms utilized by Artificial Neural Networks (ANN) provide a better generalization when compared to other classification algorithms. The generalization in ANN is influenced by three critical factors: learning rule, network architecture and training set. The ANN algorithm discussed in this paper is Cascade-Correlation, which starts with a minimal network and trains the network by adding hidden nodes dynamically. The results obtained from the four classes of algorithms are analyzed for accuracy, sensitivity and specificity.

Optimizing Bivalent Classifiers

Jim DeLeo, (National Institutes of Health Clinical Center), jdeleo@nih.gov

Abstract

Various ways to train and use bivalent classifiers are studied here in order to optimize the use of prevalence and misclassification cost information in classification tasks. Two things are demonstrated: (1) classifier performance is enhanced when these two factors are incorporated during training rather than during inferencing, and (2) probabilistic classification is usually more meaningful than pre-selected threshold rule classification.

Confidence-Based Cost-Sensitive Classification Decisions

Dragos D. Margineantu, (The Boeing Company), dragos.d.margineantu@boeing.com

Abstract

In the case of virtually all practical applications, classification algorithms are required to construct models that minimize a non-uniform loss function, rather than the 0/1 loss. One of the most efficient approaches to do this is to first estimate the class probabilities of the unseen instances and then to make the decision based on both the computed probabilities and the loss function. Learning models that compute accurate class probability estimates is - in general - known to be a difficult task. As a result, large research efforts are made to improve the accuracy of the estimates computed by different algorithms. This paper presents a novel approach to learning classification models for making cost-sensitive decisions by addressing the problem of minimizing the

actual loss associated with the decisions rather than improving the overall quality of the probability estimates. Our approach relies on employing ensembles for estimating confidences for the learned class probabilities. The classification decisions rely on the loss function and the position of the decision boundary with respect to the estimated confidence interval. For the experimental analysis we have implemented our methods using different types of ensemble algorithms: bagging, random trees, and random forests. The confidence intervals for the probability estimates are computed based (1) on counts and (2) on the normal approximation of the estimations of the base classifiers learned by the ensemble. The results show that for some tasks, the proposed algorithms outperform some of the best probability estimation-based algorithms for cost-sensitive classification. Further analysis provides other interesting insights into learning models for cost-sensitive decision making.

A Two-Stage Nearest-Neighbor Classifier with Application to Microbial Source Tracking

Jayson D. Wilbur, (Department of Mathematical Sciences, Worcester Polytechnic Institute), jwilbur@wpi.edu

Abstract

In general, nearest-neighbor methods classify an object based on the group membership of the training observations within a certain neighborhood of the object in question. These methods share both the advantages and the disadvantages of other methods for distribution-free inference. In this talk a two-stage nearest-neighbor classifier is proposed which attempts to exploit the advantages of the (single-stage) nearest-neighbor classifier while simultaneously reducing the extent to which the classifier is overfit to the training data. This present work is motivated by the problem of microbial source tracking, which attempts to trace the source of bacterial pathogens in water resources using genetic fingerprints. Applications of the proposed methodology to real and simulated data will be presented as time permits.

Friday, 10:30 a.m. to 12:15 p.m. – Best of the IASC II (Invited Session)

Resampling Techniques in Neural Networks for Nonlinear Time Series Analysis

Michele La Rocca, (Dept. of Economics and Statistics, University of Salerno, Italy), larocca@unisa.it, and
Cira Perna, (Dept. of Economics and Statistics, University of Salerno, Italy), perna@unisa.it

Abstract

Artificial neural networks are widely accepted as flexible tools of modelling complex nonlinear and dynamic systems. They are particularly useful when the underlying process relationships are not fully understood or when the nature of the phenomenon being modelled may display chaotic properties. Training a network on a time series is not hard but the problem is the reliability of estimates for the weights to assess the ability of the trained network to generalize. Their asymptotic normal distribution has been given under quite general conditions and it can be used to construct approximated inference procedures. Alternatively, a better accuracy then standard asymptotics can be achieved by using bootstrap techniques. In this talk we focus on resampling schemes in stationary and mixing non linear processes modelled by feedforward neural networks and we investigate their use for calculating confidence intervals for predictions and critical values for statistical tests. These schemes are designed to take into account that neural network models are basically atheoretical, employed for the lack of knowledge about the functional form of the data generating process. As a consequence they are intrinsically misspecified, being an approximation of the underlying model. We provide the theoretical basis for the proposed applications and we discuss some results on simulated and real data.

Combining Ordinal Measures in Medical Research

Knut M. Wittkowski, (The Rockefeller University), kmw@rockefeller.edu

Abstract

It is rare that a single variable is sufficient to represent all relevant aspects of genetic risk, genomic activity, proteome expression, clinical response, or adverse events. Since biological systems tend to be neither linear, nor hierarchical in nature, the assumptions of traditional multivariate statistical methods based on the linear model can often not be justified on theoretical grounds. We proposes the use of u-statistics for scoring multivariate ordinal data and a family of non-parametric tests for analysis. These tests are demonstrated to be special cases of a unifying concept allowing for the analysis of partially ordered data. While much of the theoretical work was done between 1937 and 1965, the proposed generalizations became feasible only with the advent of sufficiently powerful computers. Two bioinformatics tools, are available from The Rockefeller University Web site, a downloadable spreadsheet for visualization of partial ordering, and a Web service for expl! oring profiles of genetic, genomic, and proteomic data on a multi-processor cluster. The method is demonstrated with applications ranging from animal studies over clinical studies to sports (Olympics, baseball).

Friday, 10:30 a.m. to 12:15 p.m. – Statistical and Metrological Issues in Proteomics using Time-of-flight Mass Spectrometry (Invited Session)

Exploring Bioinformatics in Serum Proteomic Analysis for Early Detection of Prostate Cancer

Bao-Ling Adam, (Medical College of Georgia), badam@mail.mcg.edu,
Yutaka Yasui, (Fred Hutchinson Cancer Research Center), yyasui@fhcrc.org,
Ziding Feng, (Fred Hutchinson Research Center), zfeng@fhcrc.org, and
O. John Semmes, (Eastern Virginia Medical School), semmesoj@evms.edu

Abstract

Our laboratory has demonstrated that SELDI protein profiling of serum coupled with an artificial intelligence data analysis algorithm can effectively differentiate prostate cancer (PCA) from benign prostate hyperplasia (BPH) and unaffected healthy men (Adam, B.L. et. al. Cancer Research 62: 3609-3614, 2002; Qu, Y. et al. Clinical Chemistry 48:10, 1835-1843, 2002). We constructed a pattern matching algorithm using Surface Enhanced Laser Desorption/Ionization (SELDI) mass spectrometry serum protein profiles from 167 PCA, 77 BPH and 82 healthy men. A sensitivity of 83

Data-Driven and Peak-Based Feature Selection in Serum Protein Mass Spectrometry

Walter S. Liggett, (National Institute of Standards and Technology), walter.liggett@nist.g
Peter E. Barker, (National Institute of Standards and Technology), peter.barker@nist.gov,
O. John Semmes, (Eastern Virginia Medical School), semmesoj@evms.edu, and
Lisa H. Cazares, (Eastern Virginia Medical School), CazareLH@EVMSMAIL.EVMS.EDU

Abstract

Chemical spectra such as protein time-of-flight mass spectra are traditionally analyzed by means of peak quantitation. There are so many different proteins in, for example, a human serum specimen, that some proteomics researchers have argued against peak quantitation as an approach to feature selection. The alternative approach is data-driven feature selection such as that which is part of functional data analysis. The choice of approach must be made in reference to a particular type of data. We consider 88 spectra from repeated measurements of the same serum standard by surface-enhanced laser desorption/ionization time-or-flight mass spectrometry. The application of functional data analysis to these data requires careful registration of the spectra. Using functional canonical correlation analysis, we identify two peaks each of which is due to at least two proteins (or protein fragments), and we show that there is a strong correlation between the intensity of a protein in one of the peaks and a protein in the other. This correlation could be due to replicate-to-replicate variation in protein fragmentation that occurs in the course of the measurement. Because the two peaks are barely recognizable as each due to two or more proteins, the result we obtain by canonical correlations could not be obtained by peak quantitation. This

result is an example of the interesting data properties that might be found through broad application of data-driven feature selection to protein mass spectra.

Bioinformatics for Clinical Proteomics: Usage and Abusage

Zhen Zhang, (Johns Hopkins University), zzhang7@jhmi.edu, and
Hong Zhang, (Armstrong Atlantic State University), hong@armstrong.edu

Abstract

Advances in high-throughput genomic and proteomic expression profiling technologies have raised the expectation for molecular diagnosis and disease classification. However, the analysis of expression data generated from clinical samples poses a unique set of challenges. In addition to the typical problem of small number of samples in a high dimensional space, data from clinical samples are noisy, biased, and often with sample labeling errors. The proper use of bioinformatics tools requires an appreciation of confounding issues in multiple domains. For example, an understanding of the disease epidemiology and current clinical practice would help to determine whether the clinical samples are representative of the target population for which the endpoint of expression analysis is related to and whether the sample inclusion/exclusion criteria is likely to create systematic biases in the measured expression data. Knowledge about the impact of sample storage, processing, and laboratory analytical procedures on the expression measurement could also be important in isolating artifacts from changes that are truly related to the disease biology. In this talk, we will use clinical proteomics for cancer biomarker discovery as an example to discuss the proper usage and possible abusage of bioinformatics tools. We will present a set of protein expression analysis tools based on the Unified Maximum Separability Analysis (UMSA) algorithm and demonstrate how they might help to alleviate the impact of some of the problems in analyzing data from clinical samples.

SVD-based Functional ANOVA For Measurement Evaluation of MALDI-TOF Mass Spectrometry

Z Q John Lu, (National Institute of Standards and Technology), john.lu@nist.gov

Abstract

Scientific advances are often driven by advents of new measurement tools. Modern measurements have certainly tended to be more high throughput, producing high resolution with faster and cheaper (automated) data collection resources. The challenge in the increasingly data-driven discovery process is to reduce the data analysis time and to make sense of the massive data in near real-time. Rigorous statistics theories for high-dimensional data, or more precisely, data sets with many variables, remain

to be developed. The needs include development of novel data exploratory and visualization tools for small sample, many variables data problems and theories for valid statistical inference such as reproducibility and process control. I will describe some of our recent progress with measurement evaluation of MALDI-TOF mass spectrometer. SVD-based functional ANOVA has been developed for functional-type experimental data of molecular mass distributions of synthetic polymers samples. It is expected that our methodology may be useful in quality evaluation of mass spectra of biopolymers such as DNA or proteins as well.

Friday, 10:30 a.m. to 12:15 p.m. – The Analysis of Streaming Data (Invited Session)

Indexing Continual Range Queries for Efficient Stream Processing

Kun-Lung Wu, (IBM Watson Research), klwu@us.ibm.com,
Shyh-Kwei Chen, (IBM Watson Research), skchen@us.ibm.com, and
Philip S. Yu, (IBM Watson Research), psyu@us.ibm.com

Abstract

A large number of continual range queries could be issued against numerical data streams, such as stock prices, sensor readings, temperatures, and others. To efficiently processing these long-running queries, only the potentially relevant queries should be evaluated against the data. We develop a virtual construct-based query indexing approach to efficiently identifying the range queries that match each data object in the streams. A set of virtual constructs, e.g., intervals in 1D space or rectangular regions in 2D space, are predefined such that it is efficient to find all the virtual constructs containing any given data object. Each virtual construct has a unique ID and an associated query ID list. The query index is built as follows. Each range query is first decomposed into one or more virtual constructs. The query ID is then inserted into the query ID lists associated with those decomposed virtual constructs. Search becomes extremely efficient. For a given data object, we first find all virtual constructs covering it. Then, we report the matched queries from the ID lists associated with the covering virtual constructs. Simulations are conducted to evaluate the effectiveness of the range query indexing approach and compare virtual constructs of different shapes and sizes.

Visual Analytics for Streaming Internet Data

Edward J. Wegman, (George Mason University), ewegman@galaxy.gmu.edu, and
Karen Kafadar, (University of Colorado), kk@math.cudenver.edu

Abstract

We have been collecting Internet IP header information for nearly 1 year. The data are streaming and at the rate of 26 terabytes per year, they data cannot be stored. Therefore techniques such as recursive algorithms and evolutionary graphics must be used. The intent is to discover unusual or exotic events among the normal traffic, which may indicate attempts at intrusion. In this talk we introduce the notion of, block recursion, and illustrate some new evolutionary graphics that allow us to do some analytical assessment.

Streaming Graphics

Leland Wilkinson, (SPSS, Inc), leland@spss.com, and
Andrew Norton, (SPSS, Inc), anorton@spss.com

Abstract

Mining data streams requires algorithms that often differ from those used on static tables. Displaying data streams in charts and graphics requires algorithms very different from those used in conventional plotting. Issues range from fusing disparate time streams to rapid display updating.

Dancer is an object-oriented class library for processing changing tables of data and rendering them to a graphics environment. We present an architecture designed to integrate and buffer data streams of up to 10,000 events per second in a way that allows real-time analytics and display of up to 20 frames per second in a 2D or 3D environment.

Friday, 10:30 a.m. to 12:15 p.m. – Cancer Classification Using Gene Expression Profiling (Invited Session)

Cancer Prediction with Robust Kernel PLS Algorithm and Gene Expression Profile

Zhenqiu Liu, (Bioinformatic Cell/ TATRC), zhenqiu@bioanalysis.org,
Dechang Chen, (Uniformed Services University of the Health Sciences), dchen@usuhs.mil, and
Jaques Reifman, (Bioinformatic Cell/TATRC), reifman@tatrc.org

Abstract

One of the main objects of analyzing cancer related gene expression data is to distinguish normal and tumor samples. Since clinical predictors are not accurate enough,

genomic measures of gene expression can be useful in providing specific treatments for individual patients. This involves interpreting complex, multivariate patterns in gene expression data and assessing how much they are able to improve clinical predictions. Therefore, automated interpretation of the gene expression may pay a crucial rule is cancer treatment. Here a new robust kernel PLS algorithm is presented and applied to the recognition of the gene expression patterns. This algorithm is a robust and nonlinear version of popular partial least square (PLS) method. We compare our algorithm with other popular classification methods such as support vector machines, boosting, and C4.5 decision trees. The test results of this algorithm with gene expression data are very promising. This study demonstrates the potential applications of this algorithm for tumor diagnosis and the identification of candidate targets for therapy.

Cancer Classification Using Informative Gene Profiles

Xue-wen Chen, (The University of Kansas), xwchen@ku.edu

Abstract

The gene expression data obtained from microarrays have shown useful in cancer classification problems.DNA microarray data have extremely high dimensionality compared to the small number of available samples. An important step in microarray-based cancer classification is to select a small number of discriminant genes expressed for cancer classifications.

In this talk, I will discuss several feature selection algorithms for cancer classification. This is followed by a novel gene selection algorithm recently developed for cancer classification. This new wrapper method applies bootstrap methods in genetic search to alleviate small sample size problems and support vector machines are used for cancer classifications. Two databases for cancer classification are considered: the colon cancer database and the leukemia database. For both datasets, the proposed method successfully identifies genes informative for cancer classification and provides very reliable cancer classification results. For the leukemia dataset, the proposed algorithm is capable of identifying a small set of genes that can reveal two clusters in ALL samples when combined with hierarchical clustering algorithms.

An Efficient Max-Dependency Algorithm For Gene Selection

Hanchuan Peng, (Lawrence Berkeley National Lab), penghanchuan@yahoo.com

Abstract

In many bioinformatics problems such as cancer classfication using microarray data, it often needs to select the most characterizing genes so that they jointly have high discriminative strength to categorize the cancer type (the target variable). We develop an

efficient algorithm based on information theory to claim that for incremental gene selection, the high-dimensional max-dependency criterion is equivalent to a combination of low-dimensional max-releveance and min-redundancy criterions. Our comprehensive experimental results show that this novel method is very effective in selecting a rather small set of genes from microarray data for phenotype classification.

Friday, 10:30 a.m. to 12:15 p.m. – Trees (Contributed Session)

Some Statistical Issues Related to Feature Detection Using Random Forests

Grant Izmirlian, (National Cancer Institute), izmirlian@nih.gov

Abstract

The random forest (RF) algorithm of Breiman and Cutler is arguably one of the best "off the shelf" classification algorithms available to date, in that with practically no tuning extremely underspecified statistical problems can be classified at expected losses near the bayes error.

In the analysis of proteomic profiling data as in most all classification problems, the target is not so much the classifier per se, but the identification of important features. Towards this end, the RF algorithm supplies a peak importance measure based upon the average decrease in correct votes that occurs when a given feature is "noised".

Via Monte Carlo study, it is shown that (1) under the null hypothesis, the normalized importance measures display non-normal fat tailed asymptotics, so that a step down procedure such as the Benjamini Hochberg False discovery rate results in observed false discovery rates that are highly inflated and (2) some implications about power and sample size are hinted at using a second Monte Carlo study generated under an alternative hypothesis containing an important peak having simple odds ratio 10 for the affected class in a balanced design.

Learning From Extremely Imbalanced Data with Random Forests

Andy Liaw, (Merck Research Labs), andy_liaw@merck.com,
Chao Chen, (UC Berkeley), chenchao@stat.berkeley.edu, and
Leo Breiman, (UC Berkeley), leo@stat.berkeley.edu

Abstract

Classification of extremely imbalanced data arises in many practical applications; e.g., fraud detection, drug discovery, direct marketing, rare disease diagnosis, etc., where

the class of interest comprise of very small fraction of the data. Usually the primary interest in learning from such data is to have high accuracy in predicting the rare (positive) class, whereas the prediction accuracy of the large (negative) class is relatively de-emphasized. Such data can pose a serious challenge to most of the classification algorithms in common use, because these algorithms, without modification, aim at maximizing the overall accuracy. Even algorithms that can accommodate class weights can break down in the extreme cases. Recent research in dealing with imbalanced data suggests that sampling techniques can be quite effective. We propose two modifications to the random forest algorithm (Breiman, 2001) for such data; one is based on an ensemble of trees where the majority class is down-sampled, and the other takes class weights into account in both tree induction and counting votes for the ensemble. We compare the performance of these two approaches with some algorithms that were proposed recently by various authors (one-sided sampling, SHRINK, SMOTE, and SMOTEboost). An empirical comparison on a few datasets shows that both approaches tend to perform better than the existing methods, but neither of these two shows a clear advantage over the other.

Tree Clustering of Gene Expression Data

Lidia Rejto, (University of Delaware), rejto@udel.edu, and
Gabor Tusnady, (Renyi Mathematical Institute, Budapest, Hungary),
tusnady@renyi.hu

Abstract

With the help of an extension of k-means clustering method we construct cluster trees of gene expressions. Most gene is located on a specific position in an edge of the tree between two cluster centers. This information should be used for further study of interrelation and expression studies of genes. The tree shows the structure of the clusters and a possible evolution connections among different clusters of genes and it helps to identify networks of genes. The method applied for several data set and the applications will be presented in the talk.

Some Light in a Dark Forest - a Closer Look at Tree Model Ensembles

Simon Urbanek, (University of Augsburg), simon.urbanek@math.uni-augsburg.de

Abstract

Tree based models are quite popular due to their versatility and interpretability. It is often possible to communicate the results to domain experts with less profound statistical knowledge. On the other hand tree-based classifiers can outperform many

competing methods if tree models are augmented by the use of whole ensembles of trees or forests. Unfortunately the interpretability of the individual models is lost in such process. In this talk we will present some methods of exploratory model analysis that can be used to extract additional information about the underlying data structure from the seemingly unpenetrable forests, allowing an assessment of the models in respect to the data analytic goals and making a compromise between prediction accuracy and interpretation possible. The methods will be illustrated on clinical datasets including genomic data.

Friday, 1:45 p.m. to 3:30 p.m. – Highlights of the SAMSI Data Mining Year (Invited)

Combinatorial Search in Data Mining

David Banks, (Duke University), banks@stat.duke.edu, and
Leanna House, (Duke University), house@stat.duke.edu

Abstract

Modern data mining requires model selection, outlier detection, dimension reduction, and many other steps that entail hard combinatorial search. This talk describes some of the work that has occurred during the SAMSI data mining year to develop and compare various search strategies. The ideas are illustrated through several applications in problems that arise in robustness, multidimensional scaling, and variable selection.

Issues in 'Real Data' Mining

Ashish Sanil, (National Institute of Statistical Sciences), ashish@niss.org

Abstract

Data sets involved in typical data mining applications are usually of poor quality – containing outliers, missing values, inconsistencies, duplicates, and so on. Initially, several of their properties (such as levels for categorical variables, ranges for real-valued ones, and inter-variable relationships) are largely unknown. They tend to be large and spread out over multiple files and tables. Therefore a significant effort in data consolidation, cleaning and exploration is required. Strategies for evaluating and coping with data quality problems, and for dealing with scalability obstacles need to be formulated. This talk will focus on these issues in the context of the testbed data sets used in the SAMSI Data Mining and Machine Learning Program.

Mixtures of Factor Analyzers: Their Place in Data Mining

Ernest Fokoue, (SAMSI), epf@samsi.info

Abstract

Factor Analysis (FA) is one of the most popular techniques for dimensionality reduction. Mixture Models (MM) on the other hand allow the construction of very flexible models through their inherent ability to capture heterogeneity in the data. The Mixture of Factor Analysers (MFA) model combines the strengths of both FA and MM by allowing a simultaneous partitioning of the input space into different clusters while performing a local dimensionality reduction in each of the subspaces. This combination of clustering and dimensionality reduction is very appealing to data mining. In this talk, I will present the advantages and the challenges of the MFA model. I will also explore some uses of the MFA model in Data Mining and Machine Learning.

Friday, 1:45 p.m. to 3:30 p.m. – Epistasis (Invited)

Gene-Gene and Gene-Environment Interactions in Genetic Case-Control Association Studies

Jurg Ott, (Rockefeller University), ott@rockefeller.edu, and
Josephine Hoh, (Yale University), Josephine.Hoh@yale.edu

Abstract

In case-control association studies, individuals are genotyped for a number of single-nucleotide polymorphism (SNP) markers. Testing for main effects amounts to finding differences in genotype frequencies between case (patients) and control individuals. Another source of genetic effects underlying complex traits are interaction effects, that is, correlations among two or more SNPs that are present in one group and absent in the other. To test for differences in interaction effects between case and control individuals, for n SNPs we propose to investigate all n(n-1)/2 possible pairs of SNPs. For each pair, we carry out a likelihood ratio test resulting in a p-value. Then we order pairs by p-value and carry out the FDR procedure (false discovery rate) to see which of the possibly large number of pairs are significant. In an application to diabetic nephropathy data with 45 SNPs (990 pairs of SNPs), 374 cases and 392 controls, two pairs of SNPs exhibited interaction differences between the two groups at the 5

Detecting Epistatic Interactions Contributing to Quantitative Traits

Rob Culverhouse, (Washington University School of Medicine), rob@ilya.wustl.edu,
Tsvika Klein, (Washington University School of Medicine), tklein@im.wustl.edu,
and
William Shannon, (Washington University School of Medicine), shannon@ilya.wustl.edu

Abstract

The restricted partition method (RPM) is a partitioning algorithm for examining multilocus genotypes as (potentially non-additive) predictors of a quantitative trait. The motivating application was to develop a robust method to examine quantitative phenotypes for epistasis (gene-gene interactions), but the method can be applied without modification to gene-environment interactions. Simulation results indicate that the method provides an efficient way to identify loci contributing epistatically to a quantitative trait, even if the loci have no single locus effects. Statistical significance can be estimated through permutation testing. An example using real data involving the metabolism of a chemotherapy drug is included for illustration. Although the examples in the paper involve 2-locus interactions, the RPM is computationally feasible for the analysis of more than two loci or factors.

Systems Biology Thought Experiments for Interpreting Epistasis Models

Jason H. Moore, (Vanderbilt University), moore@chgr.mc.vanderbilt.edu

Abstract

A goal of systems biology and human genetics is to understand how DNA sequence variations impact human health through a hierarchy of biochemical, metabolic, and physiological systems. We present here a research strategy that can be used to generate hypothetical systems biology models that are consistent with pre-defined epistatic models of disease susceptibility. The goal of these studies is to perform thought experiments about the nature of complex biological systems that are consistent with epistatic models of disease susceptibility. It is anticipated that the utility of this approach will be the generation of biological hypotheses that can then be tested using experimental systems.

Friday, 1:45 p.m. to 3:30 p.m. – Analysis of Very Large Data Sets (Invited)

Having It All

Allan Wilks, (AT&T Labs - Research), allan@research.att.com

Abstract

A major problem in building a large transaction database is determining the completeness of the data. I will describe my experience in this area with reference to an 8 TB database at AT&T that has about 200 billion records in it, arriving at the rate of about 350 million records per day. The data is collected from about 500 sources that have a wide range of reporting frequency, data volume and reliability. Do we have a complete list of possible sources? Are all sources reporting? Why does a source become quiet for a period of time? Are we getting everything a source is sending? Does every record we receive make it into the database? What can we say about the completeness of the database with respect to old data? Has any corruption crept in? I will deal with these and other questions of completeness.

Interactive Graphics for Large Data Sets: There is More to it Than Meets the Eye

Antony Unwin, (Augsburg University), unwin@math.uni-augsburg.de

Abstract

Overplotting, insufficient screen resolution, inadequate detail in displays, lack of screen real estate: just some of the difficulties in visualising large data sets which obscure information in the data. Extracting what is underneath the surface takes more than basic interactive tools and will require more sophisticated software than is currently available. This paper describes the issues, illustrates some solutions and tentatively outlines future research directions.

Alternatives to Mixture Modeling in High Dimensions

David Scott, (Rice University), scottdw@rice.edu

Abstract

Clustering p-dimensional data by fitting a mixture of K normals has enjoyed renewed interest (for example, see Splus function ,mclust,). However, the number of parameters for the model grows rapidly with dimension p. For example, even if all the covariance matrices are assumed to be equal, the number of parameters is (K-1) + K*p + p(p+1)/2 for the weights, means and covariance matrix. At ACAS in 2001, Scott introduced the partial mixture component algorithm which fits only one component of the mixture model at a time. This algorithm requires only 1 + p + p*(p+1)/2 parameters for the weight, mean vector, and covariance matrix. In this talk, we introduce a new algorithm which attempts to find the ,best, line through individual clusters. This

model requires only 2*p parameters. That is, the new algorithm is linear rather than quadratic in p. By repeatedly reinitializing the search algorithm, many clusters may be identified. Intuitively, the line found is approximately the largest eigenvector of the local covariance matrix. The GGobi visualization program will be used to illustrate the success of this algorithm on real and simulated data.

Friday, 1:45 p.m. to 3:30 p.m. – Genetic and Biochemical Networks: Methods and Empirical Models (Invited)

Limitations of Statistical Learning from Gene Expression Data

Tianjiao Chu, (Institute for Human and Machine Cognition, University of West Florida), tchu@andrew.cmu.edu, and
Clark Glymour, (Institute for Human and Machine Cognition, University of West Florida), cg09@andrew.cmu.edu

Abstract

Current technologies for measuring gene expression levels, such as microarray and SAGE, measure the summed expression levels of the genes from a large aggregate of cells, rather than the expression levels of the genes in an individual cell. This paper discusses, from the statistical point of view, what we could learn, both in principle and in practice, from the microarray and SAGE gene expression level data. We show that, when the summed gene expression levels are measured from a large number of cells, the conditional independence relations among the summed gene expression levels are essentially determined by the correlation matrix among the gene expression levels, and are very unlikely to be the same as the conditional independence relations among the expression levels of the genes in an individual cell. This suggests that any algorithm for learning the gene regulatory network based on the conditional independence relations among the expression levels of the genes would not work with the data generated by the current technologies. Furthermore, we show that, in practice, we probably could not even get an accurate estimation of the correlation matrix of the gene expression levels, for the number of experiments required to estimate the correlation matrix is too large to be feasible. Therefore, the only piece of information we can learn reliably from the current gene expression level data is the expected gene expression levels.

Generating Constraints on the Topology of Genetic Networks Using Expression Data: a Combinatorial Approach

Nicholas J. I. Lewin-Koh, (Lilly Systems Biology Pte Ltd), nikko@lilly.com, and
Christopher Taylor, (Lilly Systems Biology Pte Ltd), taylor_christopher@lilly.com

Abstract

The promise of the new field of Systems Biology is that the networks of interactions between the genes and proteins in a cell will be elucidated via techniques from the quantitative sciences. As the field matures, researchers have found that reverse engineering of complete biological networks from expression data is not feasible given the current experimental techniques and data available. In this work we investigate the extent to which information on network topology can be generated by the analysis of gene expression data. Our approach is based on factorial experimental design with a functional response (e.g. gene expression over time). By combining the combinatorics of the design with the clustering structure in the data, we derive constraints on the topology of an underlying network. Our constraint rules are in the form of parent child relations of genes, based on the information content of the cluster structure of each gene in the design space. The assumption is that a parent of a node in the network topology has to contain all the design configurations of its children in its clusters. We will present preliminary results from simulated and real data.

Transient Response of Steady-State Yeast Cells to Small Perturbations

Michal Ronen, (Stanford), mronen@stanford.edu, and
David Botstein, (Princeton), botstein@princeton.edu

Abstract

An experimental design aimed at revealing the structure and dynamics of gene transcriptional networks was devised, in which yeast cells growing at steady state are exposed briefly to a different environment, and the transcriptional response is monitored over time. This design was implemented in a chemostat in which yeast cells growing in limiting galactose were pulsed with three different concentrations of glucose. All three pulses showed gene expression responses. Consistent with the idea of different regulatory mechanisms, three main kinetic behaviors of gene expression were revealed in these experiments, 1. A set of genes (including genes involved in glucose uptake, glucose metabolism, and galactose utilization) showed a correlation between the size of the pulse and maximal levels of expression, but no relationship between the size of the pulse and the dynamics of gene expression. 2. A second set of genes (including tricarboxylic acid cycle (TCA) genes) showed a relationship between the size of the pulse and the timing of the gene expression changes. 3. The last set of genes showed changes that were similar for all three pulses. Possible regulation circuits to account for these dynamics will be proposed.

Friday, 4:00 p.m. to 5:45 p.m. – Visualization and Analysis of Text/Web Data (Invited)

Efficient Visualization of Large Text Corpora

Marko Grobelnik, (J.Stefan Institute), Marko.Grobelnik@ijs.si, and
Dunja Mladenic, (J.Stefan Institute), Dunja.Mladenic@ijs.si

Abstract

Visualization is one of the important ways on how to deal with large amounts of textual data. Most frequent application of text visualization techniques is particular in cases when one needs to understand or to explain the structure and nature of large quantity of typically unlabeled and poorly structured textual data in the form of documents.

The usual approach when dealing with text for visualization is first to transform the text data into some form of high dimensional data and in the second step to carry out some kind of dimensionality reduction down to two or three dimensions that allows to graphically visualize the data. There are several (but not too many) approaches and techniques offering different insights into the text data like: showing similarity structure of documents in the corpora (e.g. WebSOM, ThemeScape), showing time line or topic development through time in the corpora (e.g. ThemeRiver), showing frequent words and phrases relationships between them (Pajek), etc.

One of the most important issues when dealing with visualization techniques is scalability of the approach to enable processing of very large amounts of the data. Our contributions are two procedures for text visualization working in linear time and space complexity.

The first procedure is a combination of the K-Means clustering procedure and a technique for nice graph drawing. The idea is first to build certain number of document clusters (with K-Means procedure), which are in the second step transformed into the graph structure where more similar clusters are connected and bound more tightly. The third step performs one sort of multidimensional scaling procedure by aesthetically drawing of the graph. Each node in the graph represents the set of similar documents represented by the most relevant and distinguishing keywords denoting the topic of the documents.

The second procedure performs hierarchical K-Means clustering procedure producing a hierarchy of document clusters. In the next step the hierarchy is drawn into the two-dimensional area split accordingly to the hierarchy splits. Like in the first approach, each cluster (group of documents) in the hierarchy is represented by the set of the most relevant keywords.

We have used the proposed approach on the number of examples visualizing e.g. Reuters text corpora (over 800k documents) and various web-sites.

Visual Text Mining with TRUST and Starlight

Anne Kao, (Boeing), anne.kao@boeing,
John Risch, (Pacific Northwest National Lab), john.risch@pnl.gov,
Steve Poteet, (Boeing), stephen.r.poteet@boeing.com, and
Jason Wu, (Boeing), jason.wu@boeing.com

Abstract

It is plainly evident that the availability of potentially useful text-based information greatly surpasses our capacity to effectively find and utilize it in a timely manner. As a response to this situation, the field of Text Mining has recently emerged as a distinct subdiscipline within the field of Data Mining. The complex and often ambiguous nature of unstructured text, however, poses unique and demanding challenges relative to those of ,classical, data mining. Problems of polysemy, metaphor, and metonymy pose difficulties for many text mining algorithms. We feel that, perhaps even more so than in ,traditional, data mining, visualization techniques can provide a valuable adjunct to text mining algorithms. In particular, we argue that visualization technologies can and should play a key role in the steering and interpretation of text mining processes. To support this added power, text mining has to be able to provide a representation not only truthful to and expressive of the information content, but also able to support visualization. We describe our recent work in TRUST and Starlight in this regard related to improving identification and interpretation of topical features associated with large document collections.

Friday, 4:00 p.m. to 5:45 p.m. – Protein Folding (Invited)

Statistical and Computational Issues in AB Initio Protein Structure Prediction

Ingo Ruczinski, (Johns Hopkins University), ingo@jhu.edu

Abstract

The prediction of protein tertiary structure from its amino acid sequence is one of the most important unsolved problems in molecular biology. We discuss the types of structure prediction, and give an overview of ,ab initio, (or ,de novo,) approaches, in particular our Rosetta algorithm which generates protein structures from fragment libraries using simulated annealing. We will describe the basics of the folding algorithm, and discuss some of the statistical and computational issues related to ab initio structure prediction. Time permitting, we will also touch on some related topics such as protein folding kinetics.

Finding the Protein-Protein Interface via Docking Calculations

Jeffrey J. Gray, (Chemical & Biomolecular Engineering,Johns Hopkins University), jgray@jhu.edu

Abstract

The protein docking problem, that is, the task of assembling two separate protein components into their biologically relevant complex structure, is important for several reasons. First, it is of extreme relevance to cellular biology, where function is accomplished by proteins interacting with themselves and with other molecular components. Second, the protein docking problem presents a fundamental test of our understanding of the energetics of macromolecular interactions, as the native complex structure is almost certainly at a global free energy minimum. Finally, an important post-genomic goal is the characterization of the structures of protein-protein complexes, and computational tools offer an inexpensive means to carry out large-scale studies. Here, I will discuss our latest methods to predict protein-protein complexes from the coordinates of the unbound monomer components. The method employs a low-resolution rigid-body Monte Carlo search followed by simultaneous optimization of backbone displacement and side-chain conformations using Monte Carlo minimization. Up to 105 independent simulations are carried out, and the resulting decoys are ranked using an energy function dominated by van der Waals interactions, an implicit solvation model, and an orientation-dependent hydrogen bonding potential. Top-ranking decoys are clustered hierarchically to select the final predictions. The talk will include both discussions of the derivation of empirical scoring functions from statistical analysis of known protein structural data and details of the computational implementation of the multi-scale algorithm in parallel on clusters of up to 60 Linux processors. The success of the algorithm was tested in an international blind challenge, the Critical Assessment of PRedicted Interactions (CAPRI). Recent notable predictions by our algorithm include one of the two best structures of the laminin-nidogen complex (T08) and a correct structure for the complex of cohesin and dockerin produced using a homology model for the starting structure of dockerin (T11). The docking source code and decoy sets (for the development of scoring functions) are available at http://graylab.jhu.edu.

Five Hierarchical Levels of Sequence-Structure Correlation in Proteins

Chris Bystroff, (Department of Biology, Rensselaer Polytechnic Institute), bystrc@rpi.edu

Abstract

A heirarchical series of statistical models have been developed for sequence-structure correlations in proteins at five levels of structural complexity: (1) short motifs, (2)

extended motifs, (3) non-local pairs of motifs, (4) three dimensional arrangements of multiple motifs, and (5) global structural homology. Here we review the statistical models, including sequence profiles, hidden Markov models and interaction potentials, for the first four levels of structural detail. The I-sites Library models local structure motifs. HMMSTR in a hidden Markov model for extended motifs. HMMSTR-CM is a model for pairwise interactions between motifs. And SCALI-HMM is a set of hidden Markov models for spatial arrangements of motifs. Sequence models for global structural homology include Pfam and SUPERFAMILY, but these will not be discussed further. Recurrent themes in the database may indicate energetically preferred states. Although the database is a fixed size, we still see recurrence as we model increasingly larger pieces of protein chain. The absence of a sparse data problem in our hierarchy of statistical models may be explained by the absence of a combinatorial explosion in the physical folding for proteins.

Friday, 4:00 p.m. to 5:45 p.m. – West Nile (Invited)

The Anatomy of a Bioevent: West Nile Virus in Washington, DC

James M. Wilson, V, MD, (ISIS Center, Georgetown University), IceAxe5@aol.com

Abstract

In late 1999, West Nile fever, which is caused by a mosquito-vectored virus, was identified in New York City during an epidemic involving 62 human cases and 7 fatalities. This was the first documentation of West Nile virus in the Western Hemisphere. It has been suggested the virus was transferred by accident via an infected human passenger arriving by airflight from the Middle East. The virus subsequently gained ecological establishment and now has been identified throughout the eastern United States, northward to Canada and as far south as Florida, with 50

West Nile virus, Dead Birds and Human Infections. A Geostatistical Approach.

Sean C. Ahearn, (Center for Advanced Research of Spatial Information (CARSI Lab.), Hunter College-CUNY), sca@geo.hunter.cuny.edu,
Constandinos Theophilides, (Center for Advanced Research of Spatial Information (CARSI Lab.), Hunter College-CUNY), ctheo@geo.hunter.cuny.edu, and
Edward S. Binkowski, (Department of Mathematics and Statistics, Hunter College-CUNY), binkowsk@hunter.cuny.edu

Abstract

The spatio-temporal relationship between unusual sightings of dead birds and human West Nile virus infections has been observed in many studies and has been proposed as an indicator of an intense amplification cycle between birds and mosquitoes. However, to date, no single study has provided quantitative evidence that the amplification cycle occurs at the local level and that it operates within certain temporal parameters. Here, throught the use of geostatistical techniques, we present the first evidence that the localized unusual space-time correspondence of dead birds models the amplification cycle and that this cycle peaks 15-16 days prior to human onset of West Nile virus infections.

Visualization, Web-Access, and Simulation of West Nile Virus Data - From the Regional to the National Level

Juergen Symanzik, (Utah State University), symanzik@math.usu.edu,
Robert Gillies, (Utah State University), rgillies@gis.usu.edu,
Samson Gebraeb, (Utah State University), samson@gis.usu.edu,
Gopi Krishna, (Utah State University), gopikrishna@cc.usu.edu
Peter Ma, (Utah State University), peterma@gis.usu.edu, and
James M. Wilson, (ISIS Center, Georgetown University), IceAxe5@aol.com

Abstract

In this talk, we will present our efforts in visualizing the spread of the West Nile Virus at the regional (Washington, D.C.) and the national level. We will then show how these visualiztaion techniques, most importantly linked micromap plots, can be used to provide Web-based access to West Nile Virus data. We will conclude with the results of our simulation studies that show how closely limited mosquito and avian data (collected at the regional level in Washington, D.C. in 2002 and 2003) are related to West Nile Virus occurrences in humans.

Friday, 4:00 p.m. to 5:45 p.m. – Statistical Analysis of Internet Data (Invited)

Network Tomography

Bin Yu, (UC Berkeley), binyu@stat.berkeley.edu

Abstract

Our professional and personal lives now depend on the internet. The heterogeneous and largely unregulated structure of the Internet renders tasks such as dynamic routing, optimized service provision, service level verification, and detection of anomalous/malicious behavior extremely challenging. The problem is compounded by the fact that one cannot rely on the cooperation of individual servers and routers to aid in the collection of network traffic measurements vital for these tasks. In many ways, network monitoring and inference problems bear a strong resemblance to other "inverse problems" in which key aspects of a system are not directly observable. This emerging new field is called Network Tomography.

In this talk, I will first review the general problem of linear internet tomography (cf. Coates, Hero, Nowak, Yu, 2002, SP Magazine) and then conver in depth a special case: the estimation of Origin-Destination (OD) traffic matrix via link counts. OD traffic matrix information is very important for dynamic updating of routing tables for the newtorks. Our approach to the OD estimation problem relies on a Gaussian model with a power relationship between the mean and variance of OD traffic over a fixed small time interval (e.g. 5 or 10 min) (cf. Cao, Davis, Vander Wiel and Yu, 2000, J. Amer. Statist. Assoc.). Recognizing Maximum Likelihood Estimation (MLE) for solving inverse problems in internet tomography is usually computationally intractable for large networks, we use (Liang and Yu, IEEE-SP, 2003) a maximum pseudo-likelihood estimation (MPLE) approach to solve a group of internet tomography problems including the OD problem. MPLE keeps a good balance between the computational complexity and the statistical efficiency of the parameter estimation. A pseudo-expectation-maximization (EM) algorithm is developed to maximize the pseudo-log-likelihood function. Finally, we will present some recent work (Liang, Yu and Taft, 2003) using a Sprint network data set with validation to compare our approach with that of the ATT group.

Empirical Analysis of Structure in Computer Network Traffic Flows

Eric D. Kolaczyk, (Boston University), kolaczyk@math.bu.edu

Abstract

Computer network traffic arises through the superposition of origin-destination (OD) flows. Hence, a thorough understanding of OD flows is essential for modeling network traffic, and for addressing a wide variety of problems including traffic engineering, traffic matrix estimation, capacity planning, forecasting and anomaly detection. However, to date, OD flows have not been closely studied, and there is very little known about their properties.

We present results of the first empirical analysis of a complete sets of OD flow time-series, taken from two different backbone networks (Abilene and Sprint-Europe). We use Principal Component Analysis (PCA) to study the intrinsic dimensionality of OD

flows and find that this dimensionality is in fact quite small. We also illustrate how the results of PCA suggest the systematic decomposition of the structure of OD flow timeseries into three main constituents: common periodic trends, short-lived bursts, and noise. We provide insight into how the various constitutents contribute to the overall structure of OD flows and explore the extent to which this decomposition varies over time. Finally, we comment on implications of these results on standard tasks in network engineering.

This is joint work with Anukool Lakhina, Dina Papagiannaki, Mark Crovella, Christophe Diot, and Nina Taft.

Hierarchical Clustering and Network Topology Identification

Rui Castro, (Rice University), rcastro@rice.edu,
Robert Nowak, (University of Wisconsin, Madison), nowak@engr.wisc.edu, and
Mark Coates, (McGill University, Canada), coates@ece.mcgill.ca

Abstract

One of the predominant schools of thought in networking today is that monitoring and control of large scale networks is only practical at the edge of the network. The effectiveness of edge-based control can be significantly enhanced by information about the internal network state, therefore methods for inferring state information from edge-based traffic measurements are of great interest. A fundamental component of the network state is the routing topology. We will address the problem of network topology identification in this talk. This problem can be formulated as a hierarchical clustering process. We present a new method for hierarchical clustering, based on a maximum likelihood formulation. Unlike other existing clustering schemes, our method is based on a generative, tree-structured model that represents relationships between the objects to be clustered, rather than directly modeling properties of objects themselves, therefore it is particularly well suited for the network problem in question. More broadly, the generative model may not reflect actual physical mechanisms relating the objects to be clustered, but it nonetheless provides a means for dealing with errors in the similarity matrix measurements, simultaneously promoting two desirable features in clustering: intra-class similarity and inter-class dissimilarity.

Saturday, 8:00 a.m. to 9:45 a.m. – Mixture Modeling of Gene Expression Data (Invited Session)

Mixture Models in Molecular Classification

Giovanni Parmigiani, (Johns Hopkins University), gp@jhu.edu,
Elizabeth S. Garrett, (Johns Hopkins University), esg@jhu.edu, and
Ed Gabrielson, (Johns Hopkins University), egabriel@jhmi.edu

Abstract

Mixture models have found a variety of application in genomic data analysis because of the multilevel nature of the data, the conveneince of defining latent subgroups, and the possibility of fitting flexible distributional forms. This presentation includes a brief overview of selected applications of Bayesian Mixture modeling in genomic analysis and a more focused discussion of a mixture-based approach to data mining and to multi-study genomic data analysis.

Specifically, we will discuss the notion of molecular profiles based on latent categories signifying under-, over-, and baseline-expression. Following this approach we can generate results that are more easily interpretable, more easily translated into clinical tools, more robust to noise, and less platform-dependent. We implement this idea via a multi-level Bayesian mixture approach that is used both for formal inverences and exploratory visualizations. We then illustrate the application of this latent variable approach to the combined analysis of gene expression studies using different technologies.

Supervised and Unsupervised Learning Methods for Gene-Expression Data

Geoff McLachlan, (University of Queensland), gjm@maths.uq.edu.au, and
Christophe Ambroise, (University of Compiegne), ambroise@utc.fr

Abstract

Microarrays have begun to be used as diagnostic tools for clinical applications. However, microarray technology in particular, and large-scale screening approaches in general , lead to the challenging problem of learning from high-dimensional data. We consider the supervised and unsupervised classification of a tumour tissue sample on the basis of its expression signature, which is the vector containing the expression levels for very many (possibly thousands) of genes. For the supervised problem (discriminant analysis), we implement recursive feature elimination (RFE) for the selection of suitable genes for use in the support vector machine in the case of multiple classes. We also consider a model-based approach for the provision of estimates of the posterior probabilities of class membership given the expression signature. For the unsupervised problem (cluster analysis), we focus on model-based approaches. In particular, we report various modifications and extensions to the EMMIX-GENE procedure. The results are demonstrated on various microarray data sets available in the bioinformatics literature.

Clustering-Based Classification for Gene Function Prediction Using Microarray Data

Wei Pan, (University of Minnesota), weip@biostat.umn.edu, and
Guanghua Xiao, (University of Minnesota), guanghx@biostat.umn.edu

Abstract

We consider application of a clustering-based classification (CBC) algorithm to gene function prediction using microarray data. First, a comparison is made with some conventional supervised learning algorithms. Second, we consider integrating the use of protein-protein interaction data with that of microarray data. Third, we propose modified CBC algorithms using consensus clustering.

Saturday, 8:00 a.m. to 9:45 a.m. – Bioinformatics I (Contributed Session)

A Novel Method for Estimating Scale Recombination Rate from Sequence Data

Osho O. Ajayi, (University of Reading, England, UK), o.o.ajayi@reading.ac.uk

Abstract

The occurrence of recombination between different regions of a genome is of interest in medical genetics, evolutionary and other related aspects of biology for various reasons. In the population genetics context, recombination is a critical issue for analysing within-species variation or variation at the population level: by averaging out the genealogical histories over some part of a genome, recombination reduces the level of stochasticity and therefore play a practical role in evolutionary inference. Traditionally, the methods for analysing genome data and making inference assumed the absence of recombination. If this assumption is wrong, the resulting inference about the evolutionary history of gene sequences are misleading and this makes the effective detection, estimation and characterisation of the recombination rate a very important issue. The many different approaches so far proposed for estimating this parameter from sequence data have been broadly classified as parametric and non-parametric based. When the former is fully used (in its various forms) for analysing data, the derived results will usually come from characteristically heavy computational workloads and are highly efficient. However, the method suffers from the setback that it is only useful for small data sets. The later is generally easy to implement, but the plausible error that may be associated with estimates raises concern. We propose a flexible and computationally attractive approach for estimating the population recombination rate from a sample of

sequence data. Simulation results showed that the method is well calibrated with very encouraging properties.

An Efficient Algorithm for Simulating Coalescence with Recombination.

Katy L. Simonsen, (Statistics Department, Purdue University), simonsen@purdue.edu,
Dan A. Noland, (CS Department and ITaP, Purdue University), nolandda@purdue.edu, and
Chinh Le, (ITaP, Purdue University), lechinh@purdue.edu

Abstract

In population genetics, the coalescent process is an important model by which the variability of DNA sequence data can be understood. Coalescent models incorporating genetic recombination have for the last 20 years played an important role in understanding the effect of linkage on genetic variability in natural populations, both theoretically and via simulation. For example, coalescence with recombination can be used to simulate the SNP marker data used to detect association with diseases and traits in humans and other non-experimental populations. However, simulation with such models (including that of Simonsen and Churchill, 1997) has suffered from a common problem: the computational complexity (computer time and memory needed) increases exponentially with the number of genetic loci involved, and with the population size and recombination rate. Thus such simulations have been limited to small numbers of loci encompassing small regions of the genome. This motivates the development of a much more efficient computer algorithm for such simulations, whose complexity is only polynomial in the parameters. I will describe the special structure of the model that made such efficiency possible, and give some timing results to show that the desired efficiency has been achieved. This new algorithm will enable the simulation of genetic data on a genome-wide scale.

Multi-protein Complex Data Clustering for Detecting Protein Interactions and Functional Organizations

Chris Ding, (Lawrence Berkeley National Laboratory), chqding@lbl.gov,
Xiaofeng He, (Lawrence Berkeley National Laboratory), xhe@lbl.gov,
Richard Meraz, (Lawrence Berkeley National laboratory), RFMeraz@lbl.gov, and
Steve Holbrook, (Lawrence Berkeley National Laboratory), SRHolbrook@lbl.gov

Abstract

Protein Interaction Networks present a useful perspective for understanding cellular processes. Recent experiments employing high-throughput mass spectrometric characterizations have resulted in large datasets of physiologically relevant multi-protein

complexes. We present a unified representation of such datasets based on an underlying bipartite graph model that present an advance over existing models of the network. This representation automatically generate protein - protein interaction network and also the protein complex - protein complex association network. Our unified representation allows for weighting of connections between proteins shared in more than one complex as well as addressing the higher level of organization that occurs when the network is viewed as consisting of protein complexes that share components. This representation also allows for the application of the rigorous spectral graph clustering algorithm for the determination of relevant protein modules in the networks. Statistically significant annotations of clusters in the protein-protein and complex-complex network using concepts from the Gene Ontology suggest that this method is also useful for detecting uncharacterized components of protein complexes or uncharacterized relationships between protein complexes.

DNAMR and DNAMRweb, Developing Easy to Use Software for DNA Microarray Data Mining.

Vladimir Kovtun and Cabrera Amaratunga, Rutgers University, kovtunv@eden.rutgers.edu

Abstract

Most existing DNA microarray data analysis software are designed for experienced Bioinformaticists. In other cases microarray software is designed for Biologists but lack the muscle of sophisticated statistical methods. The DNAMR project tries to answer this challenge by providing a web based interface DNAMR to an R package for analyzing DNA microarray data.

This talk will introduce the DNAMRweb and discuss some of the available features with examples.

Saturday, 8:00 a.m. to 9:45 a.m. – Computational Biology (Contributed Session)

Probabilistic Classification in High Dimensions for Drug Discovery

Alexander Gray, (Carnegie Mellon University), agray@cs.cmu.edu,
Paul Komarek, (Carnegie Mellon University), komarek@cs.cmu.edu,
Ting Liu, (Carnegie Mellon University), tingliu@cs.cmu.edu, and
Andrew Moore, (Carnegie Mellon University), awm@cs.cmu.edu

Abstract

Automated high-throughput drug screening constitutes a critical emerging approach in modern pharamaceutical research. The statistical task of interest is that of discriminating active versus inactive molecules given a target molecule, in order to rank potential drug candidates for further testing. Because the core problem is one of ranking, our approach concentrates on accurate estimation of unknown class probabilities, in contrast to popular non-probabilistic methods which simply estimate decision boundaries. While this motivates nonparametric density estimation, we are faced with the fact that the molecular descriptors used in practice typically contain thousands of binary features. In this paper we attempt to improve the extent to which kernel density estimation can work well in high-dimensional classification settings. We present a synthesis of techniques (SLAMDUNK: Sphere, Learn A Metric, Discriminate Using Nonisotropic Kernels) which yields favorable performance in comparison to previous published approaches to drug screening, as tested on a large proprietary pharmaceutical dataset.

Computation of the kth Nearest Neighbor Estimate of Entropy of Molecules Using Parallel Processing

E. James Harner, (West Virginia University), jharner@stat.wvu.edu,
Jun Tan, (West Virginia University), jtan@stat.wvu.edu,
Shengqiao Li, (West Virginia University and NIOSH), shli@stat.wvu.edu, and
Harshinder Singh, (West Virginia University and NIOSH), hsingh@stat.wvu.edu

Abstract

Entropy is a statistical measure of the random fluctuations in molecules and its estimation is important for investigating the stability of molecular conformations, for modeling the binding of ligands to proteins, and for studying issues relating to drug designs. Singh et al. (American Journal of Mathematical and Management Sciences, 2003, 23, 301-321) introduced a nonparametric approach for estimating entropy using the kth nearest neighbor distances between sample points which extends the first nearest neighbor approach of Kozachenko and Leonenko (Problems of Information Transmission, 1987, 23, 95-101). Entropy of a molecule depends on random fluctuations in the internal coordinates. The high dimension of the internal coordinates and a large number of observations on the coordinates cause computational challenges for computing the kth nearest neighbor distances required for obtaining the kth nearest neighbor estimate of entropy of a large molecule.

We are experimenting with computing the kth nearest neighbor distances for the sample points on a high-performance computer having a large number of parallel processors (nodes). On each processor, we use the ANN method (Arya et al. (1994), Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, 573-582) for computing kth nearest neighbor distances. ANN builds a tree structure and searches for nearest neighbors on it. In order to search for nearest neighbors concurrently using high-performance computing with many processors, we have developed a program to

duplicate the search trees on multiple processors (slave nodes) and each node computes the kth nearest neighbor distances for part of the data. When the dimension of the coordinates and the number of observations becomes large, the gain in performance becomes obvious. Using the ANN method, the whole data set needs to be loaded into the memory to build the tree. Therefore, the maximum size of the data set that ANN can handle is limited by the size of the available memory. We are also developing a program that will break the data set into parts and build tree structures on the slave nodes. To find the kth nearest neighbor of a point, the program will search all the trees concurrently and then merge all the results to get the kth nearest neighbor for the data point. The proposed program will enhance the capability of handling extremely large data sets, such as those obtained using molecular dynamics simulation, and it will improve the efficiency of computing the kth nearest neighbor distances for the sample points.

Assessment of the Relative Therapeutic Effect in Small Groups at Several Time Points: Efficacy of Mucosal and Subcutaneous Peptide Vaccines in Rhesus Macaques Exposed to SHIV

Vladimir A. Kuznetsov, (SRA International, Inc.&CIT/NIH, Bethesda MD), vk28u@nih.gov,
Vladimr S. Stepanov, (Central Economic and Mathematics Institute, RAS, Moscow, Russia), stepanov@cemi.rssi.ru,
Jay A. Berzofsky, (NCI/NIH, Bethesda MD), berzofsk@helix.nih.gov, and
Igor M. Belyakov, (NCI/NIH, Bethesda MD), belyakov@box-b.nih.gov

Abstract

Background: Due to the high cost, subject availability and ethical constraints, it is often critically important in pre-clinical and clinical studies to carry out an adequate statistical analysis of longitudinal multivariate data over several time-points in trials in several small groups. Objectives: We aim to accurately assess and develop an appropriate distribution-free longitudinal model for an estimate of the comparative treatment effects of several biological factors in several small groups even if data sets should contain outlier measurements and censored values. This approach is used to evaluate the relative efficacy of mucosal and subcutaneous polypeptide vaccines in rhesus macaques exposed to SHIV. Study design: The algorithms of the nonparametric repeated measures ANOVA models are described, programmed and assessed. Biological Results: Using nonparametric ANOVA tests, we provided statistical evaluation of the relative efficacy of mucosal and subcutaneous synthetic HIV/SIV peptide vaccines in rhesus macaques mucosally exposed to pathogenic SHIV-Ku2. We demonstrated with statistical significance that during the chronic phase of mucosal SHIV-Ku2 infection in immunized macaques, the numbers of the CD4(+) and CD8(+) cells reciprocally reflect the virus titers in plasma and both immune markers demonstrate better protection against pathogenic SHIV-Ku2 in intrarectally immunized MamuA*01 macaques. Conclusion: Despite limited data, our analysis shows better preservation of both CD4(+)

and CD8(+) cells in intrarectally immunized animals. Our analytical methodology can be applicable in comparative estimates of the different treatment-associated effects and their synergy for a variety of longitudinal data sets (laboratory data, microarray data, clinical information) in small treatment groups, even if data sets contain outlier measurements and censored values.

Identifying Differentially Expressed Proteins in 2-D DIGE Experiments

Yan Ma, (West Virginia University), yma@stat.wvu.edu, and
E. James Harner, (West Virginia University), jharner@stat.wvu.edu

Abstract

2-D gel electrophoresis is the most common technique for comparing protein abundance in different samples. Traditionally, each sample is separated on a single gel, which makes it difficult to detect differentially expressed proteins due to gel-to-gel variability. CyDye DIGE Fluor dyes allow the co-migration of more than one sample per gel. Thus an internal standard can be included along with a control and treatment in the 3-sample per gel case. This allows efficient experimental designs to be developed which essentially eliminate the gel-to-gel variability. Specialized software, e.g., Amersham's DeCyder and Nonlinear's Cross-Stain Analysis (CSA), is required to analyze the resulting stains. The ultimate goal of these analysis systems is to find differentially expressed proteins which can then be selected (by a spot picker) and sent to a MALDI-MS for identification and analysis. Unfortunately, these software systems are proprietary and methods of selecting differentially expressed proteins, e.g., using multiple testing, are unknown. However, both DeCyder and CSA allow the image-analysis data to be exported using XML and Oracle, respectively. The analyst can then intervene to do their own analysis using the import facilities of R. This allows the statistician to use state-of-the-art procedures, e.g., random forests and modern multiple testing algorithms, for identifying differentially expressed proteins. This paper will also look at issues relating to the identification, quantification, and matching of 2-D gel protein spots, i.e., the underlying image analysis.

Saturday, 8:00 a.m. to 9:45 a.m. – Large Datasets (Contributed Session)

Single-pass, Low-storage Methods for Massive Streaming Datasets with Applications to Multivariate Density Estimation

James P. McDermott, (Bristol-Myers Squibb), james.mcdermott@bms.com, and
Dennis K. J. Lin, (Pennsylvania State University), lin@chao.smeal.psu.edu

Abstract

We propose a single-pass, low-storage sequential method for the execution of multivariate density estimation for massive streaming datasets via convex hull peeling. This new method is shown to vastly reduce the computation time required for the existing convex hull peeling algorithm from $O(n^2)$ to $O(n)$. Further, the proposed method uses very low storage as compared to the existing method. We demonstrate the accuracy and reduced computation time required of the porposed method by comparing to the existing convex hull peeling method through simulation studies and a real life example

Fitting Large-Scale Spatial Models with Applications to Microarray Data Analysis

Stephan R. Sain, (CU-Denver), ssain@math.cudenver.edu, and
Reinhard Furrer, (NCAR), furrer@ucar.edu

Abstract

Many problems in the environmental and biological sciences involve the analysis of large quantities of data. Further, the data in these problems are often subject to various types of structure and, in particular, spatial dependence. Traditional model fitting often fails due to the size of the data sets since it is difficult to not only specify but to also compute with the full covariance matrix. For example a single microarray can include over 400K individual observations. We propose using a very general type of mixed model that has a random spatial component. Recognizing that spatial covariance matrices often exhibit a large number of zero or near-zero entries, covariance tapering is used to force these entries to zero. Then, using the sparse nature of such matrices and a new computational approach for computing the Cholesky decomposition, backfitting is used to estimate the fixed and random model parameters. Results will be demonstrated on a experiment using microarrays to build a profile of differentially expressed genes relating to cerebral vascular malformations, an important cause of hemorrhagic stroke and seizures.

Parallelizing the Computation of Spatial Covariance in Large Spatial Data Sets

James A. Shine, (US Army Topographic Engineering Center), jshine@tec.army.mil

Abstract

Analysis of spatial covariance structure in large spatial data sets produces results useful in optimal sampling design, compression capability, and estimation and classification

of unknown data points. However, complete analysis of such data sets is a computational challenge. The main loops of such a computation compute sums independently in each row and/or column and seem to offer near-ideal potential for parallel processing. This paper investigates the improvement in time using MPI parallelization on two Cray supercomputers and on a Linux cluster. Extensive tests on multiple images and all of these computational architectures confirm that spatial covariance computation time can be reduced by at least 2 orders of magnitude using this parallelization. These results support the thesis that real-time or near-real-time computation of spatial covariance information is an achievable goal, even for large spatial data sets.

Computationally Efficient Identification of Outliers in Large Data Sets

Mark Werner, (Oakland University), werner@oakland.edu

Abstract

We present a computationally fast procedure for identifying outliers, suitable for use in large data sets. This procedure uses a modification of Tukey's biweight function to obtain robust location and scale estimates and accordingly, robust Mahalanobis distances (RMD) for each observation. We estimate the density of these RMD's and determine a final rejection point from the empirical density function; points are then classified as outliers if their RMD is sufficiently large. Since no assumptions are made regarding the data (such as normality), this procedure demonstrates a high degree of accuracy on a wide variety of data sets, including skewed and correlated data. It is computationally efficient and is capable of rapidly identifying outliers in large, high-dimensional data sets. We also examine the influence function of the robust estimator defined by the first half of this algorithm and compute its asymptotic robustness properties. These properties are compared to other well-known estimators for a deeper understanding of robust estimation, which does not necessarily have to be performed in conjunction with outlier identification.

Saturday, 10:30 a.m. to 12:15 p.m. – Towards Understanding and Analyzing Proteomics Data (Invited Session)

Bayesian Methods for Proteomics with Feature Selection

Marina Vannucci, (Texas A&M University), mvannucci@stat.tamu.edu,
Mahlet G. Tadesse, (Texas A&M University), mtadesse@stat.tamu.edu, and
Jeffrey S. Morris, (MD Anderson Cancer Center), jeffmo@mdanderson.org

Abstract

In this talk Bayesian selection methods will be used to extract important features of proteomics data, incorporating dimension reduction wavelet techniques. We will focus on mass spectrometry data where the identification of peaks related to a specific outcome, for example peaks that discriminate samples or that predict a clinical response, is of interest. Certain preprocessing steps must be performed before analyzing proteomic spectra, such as denoising, removal of baseline, normalization and calibration of the samples. We will use wavelet reduction techniques along with Bayesian methods to select features. We will illustrate the methodology on a classification problem using proteomics data from a lung cancer diagnosis study.

The Analysis of MALDI-TOF Proteomic Spectra from Serum Samples - A Case Study

Keith Baggerly, (MD Anderson Cancer Center), kabagg@mdanderson.org

Abstract

Just as microarrays allow us to measure the relative RNA expression levels of thousands of genes at once, mass spectrometry profiles can provide quick summaries of the expression levels of hundreds of proteins. Using spectra derived from easily available biological samples such as serum or urine, we hope to identify proteins linked with a difference of interest such as the presence or absence of cancer. In this talk, we will briefly introduce one of the more common mass sprectrometry techniques, matrix-assisted laser desorption and ionization/time of flight (MALDI-TOF), and describe a case study using this technique. While we do uncover some structure of interest, aspects of the data clearly illustrate the need for careful data cleaning and preprocessing to ensure that the structure found is due to biology.

Nonparametric Approaches to the Classification of Proteomic Profiles

Kim-Anh Do, (U. T. M.D. Anderson Cancer Center), kim@mdanderson.org,
Peter Mueller, (U. T. M. D. Anderson Cancer Center), pm@odin.mdacc.tmc.edu,
Sijin Wen, (U. T. M. D. Anderson Cancer Center), sijin@odin.mdacc.tmc.edu, and
Raj Bandyopadhya, (Rice University), rajb@odin.mdacc.tmc.edu

Abstract

There has been much recent interest in using patterns in proteomic mass spectra to distinguish cancers from normals, between different stages of disease development, or between experimental conditions (such as different treatment arms).

We discuss two nonparametric methods in addressing this problem:

1. A functional data approach based on principal component analysis enhanced by the use of smoothing;
2. A model-based Bayesian inference approach by formulating the desired inference about protein expression as a density estimation problem where a proteomic profile is represented as a mixture of Beta distributions.

The performance of these methods will be assessed via an application data set of serum MALDI spectra from lung cancer and normal subjects.

Saturday, 10:30 a.m. to 12:15 p.m. – Bioinformatics II (Contributed Session)

Genome Phylogenetic Analysis Based on Extended Gene Contents

Hongmei Zhang, (University of West Flordia), hzhang@uwf.edu, and
Xun Gu, (Iowa State University), xgu@iastate.edu

Abstract

This talk is to present a methodology to estimate genome distances. We formulate the genome model of gene family evolution, under the framework of birth-death stochastic process, and then define a new additive genome distance. Because of the insufficiency of the gene-content data, we develop a novel approach to estimating the genome distance, based on the extended gene content, i.e., in a genome the status of a gene family could be either absence, presence as single-copy, or presence as duplicates. Simulations show that the new method for tree-making is efficient, consistent, and fairly robust. We apply it to 13 microbial genomes. The result supports the concept of universal tree of life.

Characterization and Re-Annotation of Common Genes Found in Complete Chloroplast Genomes

Beatrice Kilel, (George Mason University), bkilel@gmu.edu

Abstract

The recent upsurge in complete chloroplast genomes calls for a renewed focus in ensuring that the information in the public repositories is properly annotated and cited. Annotation of new genomes as well as re-annotation of existing genomes is therefore a very important step in any sequence analysis. With well annotated sequences, evolutionary studies can be performed with a lot more certainty. Phylogenetic studies are

built on the principle that if the sequences are similar, the genes to which they belong to must also be similar and tend to be functionally linked. This study focused on the re-annotation and characterization of genes found in 31 currently complete chloroplast genomes. The results obtained from re-annotation indicate changes in the number of genes originally identified in Triticum aestivum from 18 genes which encoded proteins and 8 encoded stable RNAs to additonal 4 genes believed to encode polypeptides. Adiantum capillus-veneris sequence had a 24Phylogenetic gene characterization studies indicate 3 distinct groups of non-green algae, green algae, and terrestrial plant specie based on parsimony distances. This aggregation indicates that during evolution, functional parts of DNA or protein sequences are under selective pressure, so they tend to evolve slower and are generally more highly conserved than non-functional sequences. Any local sequence conservation may indicate biological functionality. Knowing which genes is not found in a particular branch of the evolutionary tree and inferring when rearrangements occurred has enormous implications in plant regulatory network and general metabolism. Re-annotation is very crucial in order to obtain relevant and current citations, so more conclusive comparative studies can be achieved. As the field of phylogenomics continues to expand, the assignment of gene function in close juxtaposition with the species in the evolutionary tree is poised to open up new information on how genomes have retained their structure due to minimal mutations, duplication, and any major rearrangement over the evolutionary distance. The significance of the results obtained from this study is to provide a better understanding of genome rearrangements, location of homologous genes in the genomes that have not yet been sequenced, and thus investigate mechanisms of genome evolution.

Data Analysis and Modeling of the Evolution of Proteome Complexity

Vladimir A. Kuznetsov, (SRA International, Inc.&CIT/NIH, Bethesda MD), vk28u@nih.gov

Abstract

We have shown that counting the domain-to-protein links observed in the protein and protein domain/motifs data sets and analysis of statistical distributions of these counts lead to simple probabilistic model of evolution of proteome complexity of the archael, bacterial and eukaryotic organisms [1-3]. In this work using InterPro data sets, we test the basic assumptions of our model. We conclude that the domain occurrence counts in a proteome during evolution is a random birth-death quasi-steady state process such that new domains are rarely appeared as singletons and lost at constant rates, and domains are reused and lost at rates proportional to their current use.

Modeling Dinucleotide Density Fluctuations in Genome Sequences

R. H. Baran, (Office of Naval Research and Naval Surface Warfare Center), baranr@onr.navy.

Abstract

DNA sequences exhibit significant intra-genomic fluctuations with respect to a model of conditional uniformity under which (1) the expected frequency of each dinuceotide (base step) is proportional to the product of the marginal (base) frequencies and (2) the observed number follows a Poisson distribution with that conditional mean. The empirical relation between dinucleotide relative density (DRD) and Shannon mutual information in base steps is explained by broadening the model to include local modulation of the conditional mean. This modulation takes the form of a power law as local and global DRD components are log-linearly related.

Saturday, 10:30 a.m. to 12:15 p.m. – Matrix Computations and Data Mining (Invited Session)

Principal Component and Self-aggregation Clustering

Chris Ding, (Lawrence Berkeley National Laboratory), chqding@lbl.gov

Abstract

We describe principal component based clustering algorithms with well defined clustering objective functions. For the widely used K-means clustering we proved that the continuous solution of the cluster membership indicator vector is the principal component, leading to effective implementations of K-means clustering. For spectral graph clustering, we proved that scaled principal components are the continuous solutions for cluster indicator vector for a MinMaxCut clustering objective function following a min-max clustering principle that the between-cluster associations are minimized, while the within-cluster associations are maximized. Multiple scaled PCA components form a dynamic process of self-aggregation in which data objects move towards each other to form clusters, revealing the pattern of similarity. Perturbation analysis show the inherent consistency of the clustering framework. We will apply clustering methods to DNA gene expression profiles and protein interactions.

This presention is partially based on joint research with Hongyuan Zha, Xiaofeng He, Ming Gu and Horst Simon.

Operations to Construct and Maintain a Truncated ULV Decompostion

Jesse Barlow, (The Pennsylvania State University), barlow@cse.psu.edu

Abstract

Complete orthogonal decompostions are important in producing approximations to principal components. This talk will discuss the important operations involving in constructing and maintaining this approxmation. These operations include rank one updates, downdates, and refinement. In this model, only the high order principal component approximations are stored.

Classification of Microarray Data by Two-way Gaussian Mixtures

Jia Li, (The Pennsylvania State University), jiali@stat.psu.edu

Abstract

High dimensional data raise many challenges for classification algorithms. In certain applications, the attributes, each represented by a dimension, possess cluster structures. For instance, to classify disease types, microarray gene expression data are collected from cases of the concerned diseases. The number of cases is usually in the order of hundreds. Expression levels for thousands of genes may be measured for each case. Biologists tend to believe that the behaviors of the genes fall into several groups, although how the genes are grouped is unknown in a large part. In such applications, automatic attribute clustering is desirable for a better understanding of the interaction among the various attributes as well as enhancing classification performance by reducing data complexity. In this talk, I will introduce a classification approach using two-way mixture models for the purpose of optimizing classification and attribute clustering simultaneously. The mixture structure across samples (i.e. cases) is designed to approximate the density function, in a similar manner as the mixture discriminant analysis. The mixture across the dimensions plays the role of clustering attributes and controlling model complexity, critical for treating high dimensional data. Under the general mechanism of two-way mixture model, applications of quite different natures can be handled by employing particular mixture structures and distributions. Experiments with microarray data will be presented.

Unified Multiclass Proximal Support Vector Machines

Hao Helen Zhang, (North Carolina State University), hzhang2@stat.ncsu.edu, and Yongqiang Tang, (North Carolina State University), ytang@stat.ncsu.edu

Abstract

Proximal support vector machines (PSVM) is a varant of SVM, and in theory both PSVM and SVM target on the Bayes rule, which explains their comparable classification accuracy in numerious empirical studies. However, by solving a system of linear equations the PSVM demands much less computation effort than the standard SVM which deals with a quadratic programming problem. This advantage is even more important in multicategory situations, where the standard multicategory SVM can be slow due to solving a large-scaled quadratic problem under linear constraints.

Most existing PSVM procedures for k-category classification are based on one-vs-rest scheme, where the decision rule is given by the majority vote based on multiple binary classification rules. We propose a unified multicategory proximal SVM (UMPSVM) in a regularization framework of reproducing kernel Hilbert space. The UMPSVM learns the boundary among k classes simultaneously by estimating some discriminating functions which are closely associated with conditional probabilities. It embraces two-category PSVM as a special case and implements the multi-classification Bayes rule asymptotically. In addition, compared with one-vs-rest approaches, it is more straightforward to extend the UMPSVM for nonstandard situations such as unequal misclassification costs and sampling biases.

Saturday, 10:30 a.m. to 12:15 p.m. – Computational Statistics II (Contributed Session)

On Two Sample Data Analysis

Sujung Choi, (Texas A&M University), crystal@stat.tamu.edu

Abstract

We discuss two-sample problems, specifically, how to implement unified statistical methods to help application of statistical methods which can be used for both discrete and continuous data so that we can give insights to dataset. The unified statistical tools are based on concepts of mid-distribution, design of score functions, component correlations, comparison density and exponential model. Our approach to two-sample problem is to use density estimation(comparison density estimation). Specifying the density function gives a full and natural description of the data. This approach is functional in the sense that the parameters to be estimated are probability density functions. Compared with other statistical tools for two-sample problems such as t-test or Wilcoxon rank-sum test, density estimation makes us use the data more fully, which is essential in data analysis. Also our approach using comparison density gives a unified tool for several statistical tools. We can solve problems of comparison of two samples, multi samples and goodness of fit test through comparison density concept.

Jointly Optimizing Model Complexity and Data-Processing Parameters

Jim Garrett, PhD, (Becton Dickinson), jim_garrett@bd.com

Abstract

When predictor selection is applied prior to modeling, and modeling performance is assessed by cross-validation (or most other methods), then that performance estimate will be biased. When the number of predictors outstrips the number of data cases, the bias can be severe, a phenomenon known as selection bias. Fundamentally, a model process is applied–of which fitting the model is only the last step–yet performance estimation does not encompass the entire process. Cross-validation that examines the entire process is free of selection bias, yet such cross-validation presents a challenging optimization problem. I adapt an efficient multiparameter optimization algorithm, Simultaneous Perturbation Stochastic Approximation ("SPSA"), to handle loss functions having both continuous and ordered discrete inputs. SPSA is relatively efficient, handles noisy loss functions, and is unlikely to become trapped in inferior local optima, particularly for noisy loss functions. I demonstrate how this mixed-input SPSA can jointly optimize data-processing parameters (including feature-selection) and model-complexity parameters, and also provide a type of cross-validation performance estimate that is free of selection bias.

Noncentral Generalized F Distributions with Applications to Joint Outlier Detection

Donald E. Ramirez, (University of Virginia), der@virginia.edu

Abstract

An algorithm for computing the noncentral generalized F distribution is discussed. This distribution is required to compute the power of tests of multiple outliers in linear models based on shifts of location and scale. Applications to joint outlier detection are given.

Performance Metrics for Group-Detection Algorithms

James V. White, (Alphatech, Inc.), james.white@alphatech.com,
Sam Steingold, (Alphatech, Inc.), sds@alphatech.com, and
Connie G. Fournelle, (Alphatech, Inc.), cfournel@mail.alphatech.com

Abstract

A group-detection algorithm attempts to identify groups of entities in relational data that belong to specific groups or subsystems, based on records of interactions among small subsets of the entities. For example, such algorithms may be used to detect groups (or systems) of interacting proteins in bio-networks based on multiple experiments, where each experiment attempts to identify only a small subset of the system being studied. Measurements are typically noisy because they contain extraneous entities that are not members of the groups being studied. Therefore, a statistical characterization of group-detection performance is needed. This paper discusses metrics for measuring the probabilistic performance of group-detection algorithms. The metrics may be used to compare algorithms and to assess their performance in Monte Carlo simulation studies. We show that several traditional performance metrics are deficient if the size of a group is very small compared to the size of the population of entities being considered. Moreover, a pair of classical metrics (such as sensitivity and specificity or recall and precision) must be used to track the two types of errors. To address these two issues, a new information-theoretic metric, termed *proficiency*, is introduced. Proficiency may be used to measure the performance of any detection algorithm, including classical hypothesis tests in statistics.

Monte Carlo Analysis of Univariate Robust Statistical Outlier Techniques

Mark W. Lukens, (George Mason University), `mlukens@gmu.edu`, and
James E. Gentle, (George Mason University), `jgentle@gmu.edu`

Abstract

Three techniques for univariate outlier identification are: Extreme Studentized Deviate(ESD), the Hampel identifier and the Rousseeuw identifier. The latter two are robust statistical techniques. The purpose of the paper is to determine how these outlier identification techniques perform under varying conditions. An experimental design along with two different Monte Carlo simulations provides insights into the problem. Under certain assumptions it is shown that the ESD identifier performs well with very small data contamination and the robust Hampel and Rousseuw identifiers perform better with large samples and with multiple outliers.

Index