37th Symposium on the Interface of Computing Science and Statistics 2005 and the Annual Meeting of the Classification Society of North America

(Interface/CSNA 2005)

Abstracts

St. Louis, Missouri, USA
8-12 June 2005

ISBN: 978-1-62276-716-8

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2005) by the Interface Foundation of North America
All rights reserved.

Printed by Curran Associates, Inc. (2013)

For permission requests, please contact the Interface Foundation of North America
at the address below.

Interface Foundation of North America
PO Box 7460
Fairfax Station, VA 22039

Phone: (703) 993-1212
Fax: (703) 993-1700

interface@galaxy.gmu.edu

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2634
Email: curran@proceedings.com
Web: www.proceedings.com

TABLE OF CONTENTS

KEYNOTE ABSTRACTS

Mixtures at the Interface 1
David Scott

The Practice of Cluster Analysis 1
Jon Kettenring

Bayesian Perspectives on Combining Models 1
Merlise Clyde

Bayesian Causal Inference from Observational Data 1
Siddhartha Chib

Predictive Learning via Rule Ensembles 2
Jerome Friedman

Extracting Biological Meaning from High-Dimensional Datasets 2
John Quackenbush

CSNA/INTERFACE ABSTRACTS

COLLABORATION IN DATAMINING

The Role of Collaborations in the Sapphire Scientific Data Mining Project 7
Chandrika Kamath

Data Mining Success = f (Definition, Collaboration, Value) 7
Arnold Goodman

Strategies for Visual Data Mining 7
Ed Wegman

MIXTURE MODELS

Topography of Multivariate Normal Mixtures 7
Bruce G. Lindsay, Surajit Ray

Time-Dynamic Mode Tracking and the Mean-Shift Algorithm 8
Hans-Georg Mueller, Ping-Shi Wu, Peter Hall

STATISTICS AND THE LAW

Racial Profiling and Selection Models 8
Katherine Barnes

To Tell the Truth: On the Probative Value of Polygraph Search Evidence 9
Stephen E. Fienberg

Statistical Models for Improving Biometric Authentication 9
Sinjini Mitra, Stephen E. Fieinberg, Anthony Brockwell, B. V. K. Vijaya Kumar, Marios Savvides

APPLICATIONS IN SCIENCES

Carbon Dioxide, Global Warming, and Michael Crichton's State of Fear 10
Bert W. Rust

Spatially Constrained Clustering with GeoInformatics of Hotspot Detection and Early Warning 10
G. P. Patil, Reza Modarres, Pushkar Patankar

Mining Massive Earth Science Data Sets for Large Scale Structure 10
Amy Braverman

Efficient Processing of Massive Data in Comprehensive Two-Dimensional Gas Chromatography with Mass Spectrometry Detection (GC GC-MS) 11
Nathaniel Beagley, Alan Willse, Jon H. Wahl

Robust Clustering of Positron Emission Tomography Data 11
Prasanna K. Velamuru, Rosemary A. Renaut, Hongbin Guo, Kewei Chen

CLUSTERING

A Fast Clustering Algorithm with Application to Cosmology 12
Woncheol Jang
Multiple Imputation for Cluster Analysis........ 13
Michael D. Larsen
Selecting the Number of Components in Mixtures: A Risk-Based Approach 13
Surajit Ray

MASS SPECTROMETRY BASED PROTEOMICS

Characterization of Environmental Stress Response in Mouse Lungs Via Mass Spectrometric Profiling........ 14
Michael Wagner
Standardization and De-noising Algorithms for Mass Spectra to Classify Whole-Organism Bacterial Specimens 14
Somnath Datta
Strategies for Mass Spectrometry Data Analysis in Cancer Proteomics........ 14
Dayanand Naik

DETECTING GENE-GENE INTERACTIONS

The Restricted Partition Method Used to Screen for Genetic Interactions 15
Robert Culverhouse, Brian Steinmeyer, William Shannon
Multifactor Dimensionality Reduction for Detecting Epistasis........ 15
Marylyn Ritchie
Epistasis in a Case-Control Study with 100,000 SNPs 16
Josephine Hoh

INFERENCE IN A CASE-CONTROL STUDY WITH 100,000 SNPS

Randomization Tests for Functional Data Based on Adaptive Truncation 16
Jong Soo Lee, Dennis Cox
Ternary Separation and Hierarchies 16
Robert C. Powers
On the Operating Characteristics of Some Non-parametric Methodologies for the Classification of Distributions by Tail Behavior 17
Rick Ott
Minimum Energy Clustering........ 17
Maria Rizzo
Higher-Order Density Estimation and Bump Hunting 17
Michael C. Minnotte

HIGH-DIMENSIONAL BIOMEDICAL DATA

Class-Preserving Mapping of High-Dimensional Biomedical Data: Visualization, Classification, Clustering........ 18
Ray Somorjai, Brion Dolenko
Generalized MDS for Data Visualization, Clustering, and Classification........ 18
Jeffrey Solka, David Johannsen
Deriving Meaningful Structure from Spectral Embedding and Clustering........ 19
Brandon Higgs, Jennifer Weller, Jeffrey Solka

MIXTURE MODELS

On a Flexible Information Criterion for Order Selection in Finite Mixture Models ... 19
Richard Charnigo, Ramani S. Pilla

Applying Dirichlet Process Mixture Models to Compositional Data ... 20
Marie Gantz

Functional Clustering of Temporal Microarray Data ... 20
Ping Ma

DATABASES AND COMPUTERS

Automated Generation of Metadata ... 20
Faleh Al-Shameri

A Database Design for Assessing Student Learning in an Online Course System ... 21
Jun Tan, E. James Harner

Assessing Data Interoperability In Federated Distributed Databases ... 21
Daniel N. Owunwanne

Developing Statistical COM Servers ... 22
David R. Lemmon, Joseph L. Schafer

Clustering Heterogeneously Distributed Data ... 22
Eduardo A. Socolovsky

APPLICATIONS IN MEDICINE

Improving the Sensitivity of Health Care Cost Predictions Using a Combination of Regression and Classifcation Procedures ... 23
Ogi Asparouhov

Multilevel Classifcation of Quantitative Cytology Data Using Cumulative Log-Odds Method ... 24
Jose-Miguel Yamal, Dennis Cox

Automatic Classifcation of fMRI Brain Images Using Smooth Asynchrony Maps ... 24
Svetlana Shinkareva, Hernando Ombao, Bradley Sutton

Modeling Exposures for DNA Methylation Profiles ... 24
Kimberly Siegmund

The Shrinkage Variance Hotelling T-Squared Test for Genomic Profiling Studies ... 25
Grant Izmirlian, Jian-Lun Xu

BEST OF SIAM DATA MINING CONFERENCE

Dynamic Detection, Visualization and Classification of Defect Structures in Molecular Dynamics Simulations ... 27
Srinivasan Parthasarathy, Sameep Mehta, Steve Barr, Alex Choy, Hui Yang, Raghu Machiraju, John Wilkins

Topic-Driven Clustering for Document Datasets ... 27
Ying Zhao, George Karypis

Gaussian Processes for Active Data Mining of Spatial Aggregates ... 28
Chris Bailey-Kellogg, Naren Ramakrishnan, Satish Tadepalli, Varun N. Pandey

Clustering with Model-Level Constraints ... 28
David Gondek, Shivakumar Vaithyanathan, Ashutosh Garg

MODELING ALCOHOL ABUSE

Ecology of Alcohol and Alcoholism ... 29
Ed Wegman, Yasmin Said

Modeling Alcohol Abuse and Consequences ... 29
Yasmin Said, Ed Wegman

Drinking and Public Health: The Role for Simulations and Models ... 30
William F. Wieczorek

TIME SERIES, NEURAL NETWORKS, HIERARCHICAL MODELS

Time Series Prediction Based on ARMA and GRNN Technology 30
Weimin Li, Jianwei Liu, Jiajin Le, Xiangrong Wang

Discrimination of Locally Stationary Time Series 30
Wolfgang Polonik, Gabriel Chandler

Bayesian Neural Networks and Variable Selection 31
R. Adam Molnar

Hierarchical Bayesian Models for Frequent Terms in Text 31
Edoardo M. Airoldi, William W. Cohen, Stephen E. Fienberg

MODEL AVERAGING AND ASSESSMENT

A Modest Improvement Towards Seeing the Trees in a Utopian Forest 32
Grant Izmirlian

Applying Ensemble Learning to Model Quantitative Structure-Activity Relations of Pharmaceutical Molecules 33
Christopher Tong, Vladimir Svetnik, Ting Wang, Andy Liaw

A Robust Meta-Classifcation Scheme for Cancer Detection 33
Gabriela Alexe, Gyan Bhanot, Gustavo Stolovitzky, Lilian Chiang, Jorge Lepre, Ram Ramaswamy, Babu Vengataraghavan

Comparison of Estimators of Generalization Error 34
Rory Martin, Kai Yu

On Evidence Weighted Mixture Classifcation 34
Richard Everson, Trevor Bailey, Wojtek Krzanowski, Derek Partridge

BIOINFORMATICS

Statistical Learning Tools for Analyzing Metabolomic Datasets 35
Xiaodong Lin

Learning Variable Covariances via Gradients 36
Sayan Mukherjee, Ding-Xuan Zhou

A Stepwise Structural Equation Modeling Algorithm to Reconstruct Genetic Networks 36
Grace Shieh, Ching-Yun Yu, Chung-Ming Chen, Juiling Huang

CLUSTER VALIDATION

Additive Tree Fitting of Individual Differences Through (Heuristic) Iterative Projection 36
H.-F. Koehn

Multiobjective Programming Methods for Applied Cluster Analysis 37
Michael Brusco

Stability Analysis in K-Means Clustering 37
Douglas Steinley

SELECTED IASC PAPERS

Data Reduction Using L_p Criteria 38
Jonathon Schuler, James E. Gentle

Empirical Bayes Thresholding in Gene Expression Analysis 38
Michael G. Schimek, Wolfgang Schmidt

Fitting a Cox Survival Model on High Dimensional Data 38
Hans C. Van Houwelingen

Differential Co-Expression: A New Concept in Analyzing Microarrays 39
Dennis Kostka, Rainer Spang

MODEL-BASED/GRAPH-THEORETIC METHODS

Variable Selection for Model-Based Clustering 39
Nema Dean, Adrian E. Raftery

Estimation and Selection of Normal Mixture Models Based on Spacings 40
Yong Wang

Model-Based Clustering Toolbox for MATLAB 40
Angel R. Martinez, Wendy L. Martinez

Discovering Backbone Structure in Graphs 40
Juan Lin

Graph-Theoretic Scagnostics 40
Leland Wilkinson, Anushka Anand

CLUSTERING AND CLASSIFYING TEXT

Estimating Probability Mass Functions from Very Sparse Data 41
Sanjeev P. Khudanpur

Knowledge Acquisition from Text 42
Dekang Lin

An Investigation of Text Mining Techniques for the Analysis of Abstracts 42
David Marchette

MODEL-BASED CLUSTERING AND CLASSIFICATION

Model-Based Clustering of High-Dimensional Data 42
Geoff McLachlan, Richard Bean

Quantitating Differences in Two Multivariate Distributions 43
Guenther Walther

On Potts Model Clustering and Kernel K-Means 43
Alejandro Murua

MICROARRAYS

Clustering Methods in Microarrays 44
Lidia Rejto, Gabor Tusnady

Microarray Analysis: Is an Ordered Gene List Enough? 44
Leonard B. Hearne, Eric Antoniou

Microarray Gene Selection Using Mantel Correlation with K-means 44
Bill Shannon, Brian Steinmeyer

Genetic Algorithms for Feature Selection using Mantel Correlation Scoring 45
John Grefenstette, Kevin Thompson, Brian Steinmeyer, William Shannon

Gene Selection Using Support Vector Machines with Nonconvex Penalty 45
Cheolwoo Park, Hao Helen Zhang, Jeongyoun Ahn, Xiaodong Lin

APPLICATIONS

Using Classifcation of Professionals to Assess Vocational Guidance Counselors and Their Clients 46
Olga Mitina, Vera Pchelinova, Leonov Sergey

Visual Assessment of Simple Association Models 47
Heike Hofmann

Discriminant Function Analysis in Forensic Authorship Attribution 47
Carole E. Chaski

Employing Priors for Classifying High Risk Prison Inmates 47
Jong-Ho Baek

Measuring Relationships Between Entities in Free Text 48
Amanda M. White, Antonio San Lippo, Christian Posse, Ryan E. Hohimer

COMPARING CLASSIFICATION METHODS

Classifying the Classifers 50
Li Li, James E. Gentle
Comparing Neural Networks and Other Multi-Layer Classification Methods 50
Jill McCracken, Jeremy Flantzer
Comparing Nonlinear Approaches for Classification 50
Carlos Alzola, Yasmin Said
Comparing Ensemble Approaches for Classification 50
James Shine

QUANTILE REGRESSION

Quantile Regression: Beyond the Average Man 51
Roger Koenker
Quantile Volcano Plots for Identifying Significant Genes in Microarray Data 51
Xia Li, Rob Culverhouse, Bill Shannon
Quantile Regression for Gene Expression Analysis in GeneChip Arrays 51
Huixia Wang

SPECTRAL METHODS IN DATA ANALYSIS

From Text Data Mining to Gene Expression Mining and Back Again 52
Jeff Solka, Brandon Higgs, Jeffinfer Weller
Dissimilarity Matrices and Spectral Projections 52
Elizabeth Leeds, David J. Marchette
Co-clustering of Social Network Data 52
John Rigsby, Je Rey Solka

MULTIDIMENSIONAL SCALING AND ET ALIA

Multidimensional Scaling Algorithms for Large Data Sets 53
Michael W. Trosset, Patrick J. F. Groenen
Local Multidimensional Scaling for Nonlinear Dimension Reduction 53
Lisha Chen, Andreas Buja
Maximal Data Piling in Discrimination 54
Jeongyoun Ahn, J. S. Marron
Estimating the Sparse Directions in the Effective Dimension Reduction Space 54
Zhihua Qiao, Jianhua Huang
Uncovering Curvature in Data 54
Jesse Spencer-Smith

BEST OF THE JOURNAL OF CLASSIFICATION

Looking at Different Options Within the COSA Clustering Algorithm 55
Jacqueline J. Meulman, Jerome H. Friedman
A Hierarchical Methodology for Class Detection Problems with Skewed Priors 56
Christopher K. Eveland, Diego A. Socolinsky, Carey E. Priebe, David J. Marchette
Estimating the Cluster Tree of a Density 56
Werner Stuetzle

NON-NUMERIC DATA ANALYSIS

Conditional Independence Modeling for Categorical Anomaly Detection 56
Chad Scherrer, Nathaniel Beagley
Usage-Based Evolution of Visual Analysis Tools 57
Elizabeth Hetzler, Stuart Rose, Dennis McQuerry, Pat Medvick

Scenario Analysis for Homeland Security 57
Olga Anna Kuchar, George Chin Jr., Paul Whitney, Katherine Johnson, Mary Powers

DISCRIMINATION AND DATA MINING

Choosing Weights for Nearest-Neighbor Classification with Linear Programming 58
Samuel E. Buttrey, W. Matthew Carlyle
Classifcation Via Interpoint Distance Profiles 58
Jayson D. Wilbur
Polya Tree Priors for Classification Error Distributions 58
Andrew Neath
SIP Load Test Automation based on Data Mining Algorithms 59
Arta Doci

CLUSTERING METHODS AND APPLICATIONS

Imputation-Free Robust Clustering Using Soft Constraints 59
Nan Lin
Applications of Parametric and Nonparametric Bayesian Predictive Clustering 59
Fernando A. Quintana
Propensity Scoring and the LATE Distribution from Unsupervised Clustering 60
Nicholas Lewin-Koh, Robert Obenchain
Profiling Price Dynamics in Online Auctions Using Curve Clustering 60
Wolfgang Jank, Galit Shmueli
Clustering for Measurement Error in Expenditure Survey Data 60
John Dixon

COMPUTATIONAL BIOLOGY

Spike and Slab Gene Selection for Multigroup Microarray Data 61
J. Sunil Rao
Bayesian Infinite Mixture Model-Based Clustering of Functional Genomics Data 61
Mario Medvedovic
Classification and Regression-Based Approaches to Protein Structure Prediction 61
Jarek Meller, Rafal Adamczak, Michael Wagner, Aleksey Porollo

GRAPH-THEORETIC PATTERN RECOGNITION

Juggling: Ensembles of Class Cover Classifiers 62
Jason DeVinney, David Marchette, Carey Priebe
Geometry of Learning: from Graphs to Continuous Spaces 62
Mikhail Belkin
Local Intrinsic Dimension Estimation with kNNGs 62
Alfred Hero, Jose Costa

ALGORITHMS

Parallel Computation of the kth Nearest Neighbor Estimate of the Entropy of Molecules Using Circular Distances 63
E. James Harner, Jun Tan, Shengqiao Li
Characterizing the Solution Path of Multicategory Support Vector Machines 63
Zhenhuan Cui, Yoonkyung Lee
Accelerating Linearly-Convergent Algorithms 63
Tim Hesterberg
Alternative Visualization of Andrews' Curves 64
Wendy L. Martinez, Angel R. Martinez

Optimal Linear Combination of Longitudinal Markers for Disease Classification 64
Ming Ji

APPLICATIONS IN BIOLOGY

Multivariate Regression Tree: Classification of Bird Assemblages Based on Their Habitat Characteristics 64
Marie-Helene Ouellette, Jean-Luc Desgranges, Pierre Legendre, Daniel Borcard

Global Classification of (Plant) Proteins across Multiple Species 65
Naomi S. Altman, Kerr Wall, Jim Leebens-Mack, Victor Albert

Phyloinformatics of Genes in Complete Chloroplast Genomes 65
Beatrice Kilel

Visualizing Primate Evolution: Reification of a Statistical Model 66
F. James Rohlf, Nina Amenta, Eric Delson, David F. Wiley, Will Harcourt-Smith, Steve Frost, Alfred L. Rosenberger, Dan A. Alcantara, Lissa Tallman

ENVIRONMENTAL APPLICATIONS

Statistical Modeling and Evaluation of Microbial Source Tracking Data from rep-PCR DNA Fingerprints 68
Luis Tenorio, Junko Munakata-Marr, John Albert

Simulation-Based Detection of Water-Borne Bacterial Contamination 68
Brian West

Scaling by Reference Conditions for Ecological Assessment 68
Samantha Bates Prins, Eric P. Smith

MODEL BUILDING: MIXTURES & BIOINFORMATICS

Analysis of Semiparametric Mixture Models, with Application to QTL Analysis 69
Jason Fine

A Mixture Model Approach to Multiple Hypothesis Testing 69
Geoff McLachlan, Geo McLachlan, Liat Ben-Tovim Jones

Searching High-Dimensional Spaces for Parsimonious Mixture Models 70
R. Pilla, Catherine Loader

TREES AND NEURAL NETWORKS

Gaussian Process Trees 70
Robert B. Gramacy, Herbert K. H. Lee

Making Use of Small Samples for Classification 71
Russell Steele, Matthew Taddy, Inti Zlobec, Nilima Nigam

Default Bayesian Neural Network Classification 71
Herbie Lee

DATA MINING

Learning Classifiers Under a Limited Budget for Acquiring Training Data 71
Balaji Krishnapuram

Nonparametric Statistics and Data Mining 72
Tamraparni Dasu

Mining Temporal Patterns 73
Feng Liang

DEVELOPMENTS IN BIOINFORMATICS

Selecting an Appropriate Clustering Algorithm for Analyzing Microarray Data 73
Susmita Datta

Estimating Network Topology and Latent Factors in A Gene Network Based on Independent Component Analysis 74
Wei-Fuh Wang, Chung-Ming Chen

Statistical Methods to Construct Transcriptional Regulatory Networks Using Gene Expression and DNA Sequence Data 74
Biao Xing, Mark van der Laan

AUTHOR IDENTIFICATION

Lexical Predictors of Personality Type 75
Shlomo Argamon

Multiple Methods and the Federalist Papers 75
Ross Sowell, Diana Michalek

Author Identifcation on the Large Scale 76
Alex Genkin, D. Lewis, D. Madigan, D. Fradkin

Author Index