Abstracts from the Joint Conference of the Classification Society of North America and Interface Foundation of North America

Washington University in St. Louis, June 8-12, 2005

(Intentional blank page.)

COLLABORATION IN DATAMINING (Invited, 10:15-12:00)

The Role of Collaborations in the Sapphire Scientific Data Mining Project

Chandrika Kamath, Lawrence Livermore Lab., kamath2@llnl.gov

Abs: The Sapphire data mining project is involved in the analysis of data from observations, experiments, and computer simulations. Over the years, we have analyzed data from problems in astronomy, remote sensing, physics experiments, turbulent mixing of fluids, etc. In this talk, I will present a brief overview of some of our applications and discuss topics which are rarely, if ever, covered in data analysis texts, including interactions with domain scientist and how to build a successful collaboration.

Data Mining Success = f (Definition, Collaboration, Value)

Arnold Goodman, University of California at Irvine, agoodman@uci.edu

Abs: Data Miners tend to attack a data problem before defining the client's real needs. They usually leave a solution before discovering knowledge, enabling decisions or facilitating actions. This is adequate for exploration, but does not add value in business or science. Also, their consulting typically lacks successful collaboration in terms of genuine client commitment, team communication and quality control. We describe a complete data mining process with successful collaboration integrated into it. Then what to do and how to do it are discussed in considerable detail.

Strategies for Visual Data Mining

Ed Wegman, George Mason University, ewegman@galaxy.gmu.edu

Abs: The human visual system is a powerful tool for recognizing patterns in data displays, hence, as a tool for hypothesis generation. There are substantial issues associated with the visualization of large-scale, multidimensional data. We discuss some visualization tools including parallel coordinates, grand tour, and saturation brushing and illustrate how we use these tools for exploratory discovery. Some strategies we employ are the so-called BRUSH-TOUR and TOUR-PRUNE strategies. Recently, there has been a recognition that approaches for the analysis of streaming data need development. We also suggest some tools for streaming data.

MIXTURE MODELS (Invited, 10:15-12:00)

Topography of Multivariate Normal Mixtures

Bruce G. Lindsay, Penn State University, bgl@psu.edu
Surajit Ray, SAMSI, sray@bios.unc.edu

Abs: The class of multivariate normal mixtures provides a rich family for fitting densities in higher dimensions. It is known that if one mixes several normals with similar mean values, one typically ends up with a unimodal density. Although it is conventional to consider each component as a cluster, in this situation it seems better to interpret the various normal components as being refinements in the shape of a single cluster. Extending this, if there were two modes to the density, then we could think of the data as having two separated unimodal clusters, where the shape of the cluster was determined by the components it included. Construction of such a hierarchical clustering scheme requires understanding the modality structure of multivariate normal mixtures. We will show how one can determine the ridgelines of normal mixtures, which gives a lower dimensional way to find the modes and saddlepoints of the mixed density.

Time-Dynamic Mode Tracking and the Mean-Shift Algorithm

Hans-Georg Mueller, UC Davis, `mueller@wald.ucdavis.edu`
Ping-Shi Wu, UC Davis, `pwu@ucdavis.edu`
Peter Hall, Australian National University, `halpstat@maths.anu.edu.au`

Abs: We introduce a time-dynamic kernel type density estimate for the situation where an underlying multivariate distribution evolves with time. This is used to estimate time-dynamic modal evolution paths. The estimators involve boundary kernels for the time dimension so that the estimator is always centered at current time, and multivariate kernels for the spatial dimension of the time-evolving distribution. A time-dynamic algorithm for online mode tracking is proposed, including automatic bandwidth choices, and is implemented via a fast mean update algorithm. Simulations and animations illustrate the proposed methods, and these demonstrations are complemented with results on uniform convergence in both time and space. This is joint work with Peter Hall and Ping-Shi Wu.

STATISTICS AND THE LAW (Invited, 10:15-12:00)

Racial Profiling and Selection Models

Katherine Barnes, Washington University School of Law, `kbarnes@wulaw.wustl.edu`

Abs: This paper introduces a Bayesian model to control for selection bias. The primary innovation is that the model does not require individual-level data for the underlying population; thus, it has broader application than the standard Heckman-style selection models. Instead, the model relies upon population-level data to identify an appropriate prior. The paper provides an application in the racial profiling context, where data of stops and searches of individual motorists on the highway exist, but there is no individual-level data on the underlying population of motorists on the highway.

To Tell the Truth: On the Probative Value of Polygraph Search Evidence

Stephen E. Fienberg, Carnegie Mellon University, fienberg@stat.cmu.edu

Abs: Polygraph evidence of deception or truthfulness has largely be excluded from criminal and other legal proceedings in the U.S., beginning with the famous Fry ruling in 1923. But is such exclusion based on sound statistical reasoning, especially in the context of a screening search? In this presentation we review what we know (and don't know) about polygraph accuracy, and its use in both security screening and criminal screening contexts. In particular we draw on a recent study from the National Research Council to provide credible data on polygraph accuracy and we reconsider the probative value of both positive and negative polygraph results. We conclude that courts should continue to be skeptical of such evidence in legal settings.

Statistical Models for Improving Biometric Authentication

Sinjini Mitra, Carnegie Mellon University, smitra@stat.cmu.edu
Stephen E. Fieinberg, Carnegie Mellon University, fienberg@stat.cmu.edu
Anthony Brockwell, Carnegie Mellon University, abrock@stat.cmu.edu
B.V.K. Vijaya Kumar, Carnegie Mellon University, kumar@ece.cmu.edu
Marios Savvides, (Carnegie Mellon University), marioss@andrew.cmu.edu

Abs: The modern world has seen a rapid evolution of the technology of biometric authentication, prompted by increasing urgency to ensure system security. The need for efficient authentication systems has skyrocketed since 9/11, and the proposed inclusion of digitized photos in passports shows the importance of biometrics in homeland security today. Based on a person's essentially unique biological traits, these methods are potentially more reliable than traditional methods like PINs and ID cards. This talk will focus on establishing a firmer statistical foundation for biometric authentication systems, and present a statistical framework for evaluating the accuracy of current methods which are purely empirical in nature. We first present an existing non model-based face authentication system based on a linear filter called the Minimum Average Correlation Energy (MACE) filter, and describe how simple statistical models can evaluate its performance on large real-world databases containing diverse images. We then propose a novel mixture model-based approach in the frequency domain by exploiting the well-known significance of phase in face identification (Hayes 1982). Some classification results, inference and associated challenges are discussed. Next, a novel feature-based approach based on facial asymmetry in the frequency domain is introduced, along with classification results and comparison with results from analogous spatial domain measures. Finally, potential extension of the statistical framework to fingerprints is presented.

APPLICATIONS IN SCIENCES (Contributed, 10:15-12:00)

Carbon Dioxide, Global Warming, and Michael Crichton's *State of Fear*

Bert W. Rust, National Institute of Standards & Technology, bert.rust@nist.gov

Abs: In his recent novel, *State of Fear* (HarperCollins, 2004), Michael Crichton questioned the connection between global warming and increasing atmospheric carbon dioxide by pointing out that for 1940-1970, temperatures were decreasing while atmospheric carbon dioxide was increasing. A reason for this contradiction was given at Interface 2003 [B.W. Rust, Computing Science and Statistics, 35 (2003) 263-277] where the temperature time series was well modeled by a 64.9 year cycle superposed on an accelerating baseline. For 1940-1970, the cycle decreased more rapidly than the baseline increased. We have entered another cyclic decline, but the temperature hiatus this time will be less dramatic because the baseline has accelerated. This paper demonstrates the connections between fossil fuel emissions, atmospheric carbon dioxide concentrations, and global temperatures by simultaneously modeling their measured time series.

Spatially Constrained Clustering with GeoInformatics of Hotspot Detection and Early Warning

G.P.Patil, Penn State University, gpp@stat.psu.edu
Reza Modarres, (George Washington University), reza@gwu.edu
Pushkar Patankar, (Penn State University), pushkar@psu.edu

Abs: A declared need is around for geoinformatic surveillance statistical science and software infrastructure for spatial and spatiotemporal hotspot detection and early warning. What we particularly need is capability to detect arbitrarily shaped hotspot clusters. Our innovation employs the notion of an upper level set, and is accordingly called the upper level set scan statistic (see Patil and Taillie, 2004, Environmental and Ecological Statistics, 11, 183-197). In this presentation, we will discuss the role and use of spatially constrained clustering in identifying the candidate hotspots as connected components of upper level sets induced by the cellular surface of the response variable on the cellular tessallation of the region involved. Applications for both synoptic regions and networks of various kinds will also be discussed.

Mining Massive Earth Science Data Sets for Large Scale Structure

Amy Braverman, Jet Propulsion Laboratory, Amy.Braverman@jpl.nasa.gov

Abs: The traditional way to look for large scale structure in very large observational or model generated data sets is to examine maps of means and standard deviations of parameters of interest on a coarse spatio-temporal grid. This approach is popular because it is easy to implement and understand, but unfortunately it throws away almost all of the distributional information in the data. Moreover, maps are computed for individual parameters of interest, and therefore do not retain information about relationships among two or more parameters. In this work, we use a modified data compression algorithm to produce multivariate distribution estimates for each grid cell. The algorithms optimally mediates between data reduction and fidelity loss using information-theoretic principles. Changes in these distribution estimates over time, space and resolution reflect large scale data structure. This is the basis for a data mining algorithm that characterizes those changes using a pseudo-metric for the distance between distributions. We demonstrate using data from the Atmospheric Infrared Sounder (AIRS) on board NASA's Aqua satellite.

Efficient Processing of Massive Data in Comprehensive Two-Dimensional Gas Chromatography with Mass Spectrometry Detection (GC×GC-MS)

Nathaniel Beagley, Pacific Northwest National Laboratory, Nathaniel.Beagley@pnl.gov
Alan Willse, Pacific Northwest National Laboratory, Alan.Willse@pnl.gov
Jon H. Wahl, (Pacific Northwest National Laboratory), Jon.Wahl@pnl.gov

Abs: Modern laboratory technologies are increasingly used in high-throughput screening applications. Gas chromatography (GC) can be used to screen compounds in chemical mixtures. For complex mixtures with unknown constituents, comprehensive two-dimensional gas chromatography (GC×GC) enables high-resolution separation, potentially separating solutes that overlap in GC. To identify the compounds an additional step of mass spectrometry (MS) is used. We are applying this hyphenated technique (GC×GC-MS) to large-scale screening experiments, where the compounds are to be selected from all compounds detected in a sample, and are not limited to pre-selected target compounds. Underlying statistical problems for screening include component detection, quantification, and matching between samples. The ideal statistical model for this data is actually conceptually simpler than the ideal statistical model for GC data as the enhanced separation lessens the statistical load. The real challenge, however, will be efficient data processing. GCxGC-MS produces massive data files on the order of 1.5GB per sample which existing commercial software can often take several hours to process. And, as a typical screening experiment can have on the order of a hundred samples, incremental data mining techniques need to be utilized to solve the massive data problem. We will describe our approach to analyze GC×GC-MS data illustrated on samples before and after treatment where the goal is to find compounds affected by the treatment.

Robust Clustering of Positron Emission Tomography Data

Prasanna K. Velamuru, Arizona State University, Prasanna.Velamuru@asu.edu
Rosemary A. Renaut, Arizona State University, renaut@asu.edu
Hongbin Guo, Arizona State University, hb_guo@asu.edu
Kewei Chen, Banner Good Samaritan Medical Center, kchen@math.asu.edu

Abs: Positron Emission Tomography (PET) data, in general, is difficult to segment due to low signal to noise ratio. However, in recent years, it has been demonstrated that clustering

can be used as an important preprocessing step prior to parametric estimation from dynamic PET data. Classical clustering methods such as hierarchical clustering and k-means have been used to improve the accuracy of voxel level quantification in PET images. To obtain meaningful cluster groupings, it is necessary to perform clustering using an appropriate weighting technique, for each slice, based on the different time instants at which the data is sampled. Traditional hierarchical clustering methods require one to maintain a connectivity matrix showing the (dis)similarities between voxels. This requirement places severe computer memory constraints due to the high dimension of PET data. Prior work done by our group has shown that an application of a two-stage clustering process that combines a preclustering scheme along with classical hierarchical analysis can be used as a fast clustering alternative for dynamic PET data. In spite of the significant advances made in the application of various clustering strategies to dynamic PET data, very little work has been done related to validation of cluster results and methods in this domain. In this paper, we compute and analyze several known intra-cluster measures, inter-cluster measures and indices that are a combination of these measures. Quantification and evaluation of the clustering results with the help of these measures is important in getting an estimate of the optimal number of clusters within the data. This information is crucial since most of the commonly used clustering methods are unsupervised and it is difficult to know where to cut the hierarchical tree or how many partitions are desired. Cluster validation is also applied to compare different clustering methods with respect to efficiency and accuracy, in which accuracy is measured with respect to whether or not a clustering method that is less computationally expensive still maintains the characteristics of an expensive clustering method.

CLUSTERING (Invited, 2:15-4:00)

A Fast Clustering Algorithm with Application to Cosmology

Woncheol Jang, ISDS, Duke University, wjang@stat.duke.edu

Abs: We present a fast clustering algorithm for density contour clusters Hartigan (1975) that is a modified version of the CFF (Cuevas et al 2000) algorithm. By Hartigan's definition, clusters are the connected components of the level set $S_c = f > c$ where f is a probability density function. We use kernel density estimators and orthogonal series estimators to estimate f and modify the CFF algorithm to extract the connected components from the plug-in level set estimators. Unlike the original algorithm, our method does not require an extra smoothing parameter and can use the Fast Fourier Transform (FFT) to speed up the calculations. We show that the cosmological definition of clusters of galaxies is equivalent to density contour clusters and present an application in cosmology.

Multiple Imputation for Cluster Analysis

Michael D. Larsen, Iowa State University, larsen@iastate.edu

Abs: Cluster analysis is used to identify homogeneous groups in data. Methods have been developed to cluster binary, categorical, quantitative, and mixed data. Methods such as mixture models and latent class models are based on probability models for the data. Techniques such as k-means, hierarchical agglomeration, and monothetic analysis are defined procedurally or algorithmically without an underlying data distribution. Often some observations are missing. This poses a challenge for clustering procedures, especially for those without underlying probability distributions. Current methods handle missing data by dropping cases or variables from some or all computations or by filling-in missing values according to a prediction rule. Sometimes data are completed with a single set of imputations before clustering. In any case, the current methods do not incorporate uncertainty due to missing values into summary statements. Multiple imputation uses probability models to generate imputations from random distributions conditional on the observed data. This talk presents methods for multiple imputation for missing data in the context of cluster analysis. Thus models consistent with cluster structure produce the imputations. Models vary by type of data. Rules are developed for combining cluster analysis results from multiply imputed data sets. Methods are demonstrated on various data sets.

Selecting the Number of Components in Mixtures: A Risk-Based Approach

Surajit Ray, Statistical and Applied Mathematical Sciences Institute, sray@samsi.info

Abs: Multivariate mixture models provide a convenient method of density estimation and model based clustering as well as providing possible explanations for the actual data generation process. But the problem of choosing the number of components (g) in a statistically meaningful way is still a subject of considerable research. Available methods for estimating g include, AIC and BIC, estimating the number through nonparametric maximum likelihood, hypothesis testing and Bayesian approaches with entropy distances. We propose the idea of risk based model selection as an approach to estimating the number of components in a mixture model. The loss function is defined as the distance between the true density and the proposed mixture model, and we choose the g which has the minimum risk. Unlike the AIC and BIC, which compares between two competitive models at each stage, our risk-based method can be used to determine a global comparison between all competitive models at the same time. We will mostly deal with multivariate data, potentially with a huge number of variables; so to avoid multidimensional numerical integration in the risk calculation, the distance is chosen appropriately (depending on the model class). We will also show that by varying the tuning parameter in the distance we can analyze a data set at different levels of smoothness. Application of the risk based methods, along with comparison to other existing methods will be discussed in the results section.

MASS SPECTROMETRY BASED PROTEOMICS (Invited, 2:15-4:00)

Characterization of Environmental Stress Response in Mouse Lungs Via Mass Spectrometric Profiling

Michael Wagner, Cincinnati Children's Hospital, mwagner@cchmc.org

Abs: Mass spectrometry can be used as a tool to rapidly detect differences in the proteome of control and challenged samples. We will outline our strategy to profile mouse lung lavages of ca. 100 genetically identical mice, 50 of which have been exposed to the environmental toxin acrolein for 24 hours. Much care must be taken to ensure technical reproducibility, and our presentation will focus on computational methods to extract reliable (reproducible) information from mass spectral data on which clustering and classification methods can be based. The outcome of our analysis is a set of differentially expressed cross-validated peaks, thus facilitating the next steps in the identification of biomarkers to acrolein exposure and, eventually, the biology of the underlying pathways. (Joint work with Michael Borchers and Anne McLachlan.)

Standardization and De-noising Algorithms for Mass Spectra to Classify Whole-Organism Bacterial Specimens

Somnath Datta, University of Georgia, datta@stat.uga.edu

Abs: Application of mass spectrometry in proteomics is a breakthrough in high throughput analyses. Early applications have focused on protein expression profiles to differentiate amongst various types of tissue samples (e.g., normal vs. tumor). Here our goal is to use mass spectra to differentiate bacterial species using whole-organism samples. The raw spectra are similar to spectra of tissue samples, raising some of the same statistical issues (e.g., non-uniform baselines and higher noise associated with higher baseline), but are substantially noisier. As a result, new preprocessing procedures are required before these spectra can be used for statistical classification. In this talk, we introduce novel preprocessing steps that can be used with any mass spectra. These comprise a standardization step and a de-noising step. The noise level for each spectrum is determined using only data from that spectrum. Only spectral features that exceed a threshold defined by the noise level are subsequently used for classification. Using this approach, we trained the Random Forest program to classify 240 mass spectra into four bacterial types. The method resulted in zero prediction errors in the training samples and in two test data sets having 240 and 300 spectra, respectively.

Strategies for Mass Spectrometry Data Analysis in Cancer Proteomics

Dayanand Naik, Old Dominion University, dnaik@odu.edu

Abs: Mass spectrometry has emerged as an important tool for analyzing and characterizing large bio-molecules of varying complexity. Matrix assisted laser desorption/ionization time-of-flight (MALDI-TOF) and the recently developed surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) technologies have enabled the study of differentiable characteristics of proteins by analyzing the mass spectrometry of serum samples. Cancer proteomics involves identification and analysis of differentially expressed proteins relative to healthy tissue counterparts at different stages of disease. The process of converting mass spectra into usable data for the analysis, to distinguish between the healthy and diseased samples, is laborious and involved. In this talk, we shall review and walk through various steps of pre-processing, namely, baseline subtraction, peak identification, intensity normalization, peak alignment, and feature selection of the spectrometry data. We will illustrate some of these steps on a real life data set and use these processed data for classification and other statistical analyses. (Joint work with Michael Wagner, Alex Pothen et al.)

DETECTING GENE-GENE INTERACTIONS (Invited, 2:15-4:00)

The Restricted Partition Method Used to Screen for Genetic Interactions

Robert Culverhouse, Washington University School of Medicine, rob@ilya.wustl.edu
Brian Steinmeyer, Washington University School of Medicine, bsteinme@im.wustl.edu
William Shannon, Washington University School of Medicine, shannon@ilya.wustl.edu

Abs: The Restricted Partition Method (RPM) was developed to detect genetic interactions affecting a quantitative phenotype in a sample of unrelated individuals, even if the main effects of the covariates are small. Using data simulated for the Pharmacogenetics Research Network Analysis Workshop, the RPM was used to scan for univariate effects and 2-way interactions among 4 environmental covariates and 133 candidate SNPs. The scan detected a significant gene-environment interaction related to a qualitative phenotype and a region of high linkage disequilibrium where multiple SNPs were significantly associated with a quantitative phenotype.

Multifactor Dimensionality Reduction for Detecting Epistasis

Marylyn Ritchie, Vanderbilt University, ritchie@chgr.mc.vanderbilt.edu

Abs: In the quest for disease susceptibility genes, the reality of gene-gene interactions, or epistasis, creates difficult challenges for many current statistical approaches. In an attempt to overcome limitations with current disease gene detection methods, we have previously developed the Multifactor Dimensionality Reduction (MDR) approach. In brief, MDR is a method that reduces the dimensionality of multilocus information to identify polymorphisms associated with an increased risk of disease. This approach takes multilocus genotypes and develops a model for defining disease risk by pooling high-risk genotype combinations into one group and low-risk combinations into another group. Cross validation and permutation testing are used to identify optimal models.

We simulated data using a variety of epistasis models varying in allele frequency, heritability, and the number of interacting loci. We estimated power as the number of times that each method identified the correct functional SNPs for each model out of a set of 10 total SNPs. Using simulated data, we show that MDR has high power to detect interactions in sample sizes of 500 cases and 500 controls, in datasets with 100, 500, 1000, and 5000 SNPs. This study provides evidence that MDR is a powerful statistical approach for detecting gene-gene interactions. MDR will continue to emerge as a valuable tool in the study of the genetics of common, complex disease.

Epistasis in a Case-Control Study with 100,000 SNPs

Josephine Hoh, Yale School of Public Health, josephine.hoh@yale.edu

Abs: I will present a preliminary analysis on SNP interaction patterns in an association study of age-related macular degeneration (AMD). The analysis includes selecting pairs of SNPs that are associated with AMD and then modeling their epistatic effects.

INFERENCE AND FOUNDATIONS (Contributed, 2:15-4:00)

Randomization Tests for Functional Data Based on Adaptive Truncation

Jong Soo Lee, Rice University, jslee@stat.rice.edu
Dennis Cox, Rice University, dcox@stat.rice.edu

Abs: The hypothesis testing aspect of functional data inference has been developed only recently, and there still is a need for improvement over existing methods. Borrowing ideas of Fan (1996) and Fan and Lin (1998), we consider the method of adaptively truncated versions of Hotelling's T-squared test statistics for tests of equality of the mean functions in the two sample framework. Combined with the randomization-based procedure for obtaining a null distribution, we show that our proposed method is attractive because of its simplicity and theoretical properties. We examine numerical results and computational issues, and apply our method to image data from a cervical cancer detection study.

Ternary Separation and Hierarchies

Robert C. Powers, University of Louisville, rcpowe01@louisville.edu

Abs: If H is a hierarchy on some finite set S, then H determines a ternary relation $s(H)$ as follows: (a, b, c) belongs to $s(H)$ if and only if there exists a cluster A in H such that A contains the objects a and b but not the object and c. So H, through the cluster A, separates the objects a and b from the object c with the interpretation that a and b are more similar to each other than either is to c. A well known and useful fact is that the function s, which maps hierarchies on S to ternary separation relations on S, is injective. We consider ternary separation from a new point of view by showing that the function s

satisfies three natural algebraic properties and that these three properties are only satisfied by functions that are closely connected to s.

On the Operating Characteristics of Some Non-parametric Methodologies for the Classification of Distributions by Tail Behavior

Rick Ott, Rice University, rott@stat.rice.edu

Abs: Classical extreme value theory partitions distributions into one of three categories based on the limiting behavior of the standardized maximum. These categories are commonly perceived as containing distributions of short, medium, and long tails respectively. Many have considered the classical medium class too abundant. The last 25 years have led to alternative methods of classifying distributions by tail behavior. This presentation will review the classical classification scheme along with a few of the alternative methods, specifically the density-quantile method of Parzen (1979), the hazard function refinements to the density-quantile approach by Schuster (1984), and the residual life function method by Rojo (1996). Refinements to one or more of these classification schemes will be presented. Tests to classify data as short-, medium-, or long-tailed by the definitions of Rojo will be presented.

Minimum Energy Clustering

Maria Rizzo, Ohio University, rizzo@math.ohiou.edu

Abs: Energy statistics are based on a generalization of Newton's potential energy due to Gabor J. Szekely. In this talk the minimum energy concept is applied to the problem of hierarchical cluster analysis. The minimum energy cluster distance function determines a statistically consistent test of homogeneity, and the hierarchical clustering procedure can be viewed as an extension of Ward's minimum variance method.

Higher-Order Density Estimation and Bump Hunting

Michael C. Minnotte, Utah State University, minnotte@math.usu.edu

Abs: We investigate the effects of applying fourth-order density estimation techniques to Silverman's (1980) test of multimodality of a density. Silverman's test uses as a test statistic the "critical bandwidth," the largest bandwidth for a kernel density estimate in which the number of modes is at least the alternative-specified k (often 2). Silverman used a normal kernel for its unique bandwidth-modality monotonicity property, but Hall, Minnotte, and Zhang (2004) showed that use of other reasonable second-order kernels is possible with a little care and produces comparable results. In this study, the second-order kernel density estimate is replaced with a variety of fourth-order (reduced bias) density estimates, using higher order kernels, multiplicative corrections, transformations, variable bandwidths, and data sharpening. Computational dangers and solutions will be discussed, along with level and power evaluations and the effects of calibration.

HIGH-DIMENSIONAL BIOMEDICAL DATA (Invited, 4:15-6:00)

Class-Preserving Mapping of High-Dimensional Biomedical Data: Visualization, Classification, Clustering

Ray Somorjai, National Research Council Canada, Ray.Somorjai@nrc-cnrc.gc.ca
Brion Dolenko, National Research Council Canada, Brion.Dolenko@nrc-cnrc.gc.ca

Abs: We discuss, highlight and extend our previously introduced similarity (distance)-based projection method for high-dimensional data. This mapping is onto a special plane, called the Relative Distance Plane (RDP). Other projection methods, such as SOM, MDS, etc., try to approximately preserve all distances; the RDP mapping preserves exactly the original distances of all points to any two reference patterns, Ri, Rj. RDP mapping provides multiple viewpoints of the data, and is a discretized version of projection pursuit, but without the need for optimization. It can readily use any distance or dissimilarity measure. Because of its speed, flexibility and versatility, RDP mapping is a powerful exploratory tool that helps detect and confirm outliers, and can assess statistically whether two groups derive from the same distribution. The RDP mapping is class preserving when the samples possess class labels, and this property naturally suggests direct classification either with respect one of the reference axes, or in the various available RD planes. In fact, any standard classifier (e.g., LDA, k-nearest neighbor, SVM, etc.) can be used directly. Furthermore, the RDP is the natural setting for Ginis transvariation-probabilities-based classification method. Sets of reference axes can also be combined to create new coordinate systems in which arbitrary classifiers can be developed. The classification results on several high-dimensional biomedical datasets will be compared.

Generalized MDS for Data Visualization, Clustering, and Classification

Jeffrey Solka, Naval Surface Warfare Center, jeffrey.solka@navy.mil
David Johannsen, Naval Surface Warfare Center, david.johannsen@navy.mil

Abs: Traditionally, Multidimensional Scaling (MDS) has been performed by seeking a configuration of points in a (low-dimensional) Euclidean space, $\mathbb{R}^p$, whose interpoint distances approximate a specified dissimilarity matrix. However, if one supposes that the original data actually reside on an embedded submanifold of high codimension, then one should consider performing MDS to more general spaces than Euclidean space in order to exploit this discovered manifold structure. We have devised algorithms which approximate the metric on the embedded submanifold and which allow the determination of configurations (via Generalized MDS schemes) in compact orientable surfaces which are equipped with their compatible constant curvature metrics (e.g., the torus with the flat metric, or a genus g¿1 surface equipped with a hyperbolic metric). We will present our algorithm and demonstrate its effectiveness on several synthetic and real-world data sets.

Deriving Meaningful Structure from Spectral Embedding and Clustering

Brandon Higgs, George Mason University, bhiggs@gmu.edu
Jennifer Weller, George Mason University, jweller@gmu.edu
Jeffrey Solka, Naval Surface Warfare Center and George Mason University, jsolka@gmu.edu

Abs: The reduction of high dimensional data into meaningful low dimensional representations is often necessary to clarify important relationships and reveal inherent structure. Non-linear data structures in high dimensional space are not accurately represented by strict Euclidean distances, and as such, not optimal for conventional methods of dimension reduction. Such methods generally seek to minimize a global cost function, which tends to distort local associations and inaccurately represent the inherent connections between points. The spectrum of the Laplacian operator preserves these neighborhood geometries as it learns the data on a low-dimensional manifold. We extend previous results on image data types with an investigation of the outcome when applying the graph Laplacian and Laplace-Beltrami operators on biological data. We find that the spectral properties of the weighted graph Laplacian have particular applicability to gene expression data as judged by the ability to classify and cluster points of known disease type and biological function as well as provide a meaningful projection map.

MIXTURE MODELS (Invited, 4:15-6:00)

On a Flexible Information Criterion for Order Selection in Finite Mixture Models

Richard Charnigo, University of Kentucky, richc@ms.uky.edu
Ramani S. Pilla, Case Western Reserve University, pilla@case.edu

Abs: Mixture models provide easily-interpreted representations of the heterogeneity in physical phenomena and biological processes; yet, finite mixture models pose special challenges to statisticians, especially with regard to estimation of the order (i.e., the number of distinct mixture components). Lindsay (1983) has developed an elegant framework for nonparametric estimation of the mixing distribution (and, hence, of the order) in the absence of a structural parameter common to all mixture components. However, we demonstrate that, under fairly general conditions, incorporation of a structural parameter results in nonexistence of the semiparametric estimator (if no restriction is placed on the structural parameter) or in a degenerate semiparametric estimator (if the structural parameter is not permitted to exceed some upper bound). Thus, a different paradigm for order selection is required to accommodate the presence of a structural parameter. We propose a flexible information criterion (FLIC) by which both the order of a finite mixture model and the value of the structural parameter can be consistently estimated. The FLIC is adaptive in the sense that the strength of the penalty is determined by the configuration of the data, a feature absent from the AIC and BIC. We investigate the performance of the FLIC through simulation experiments and applications to real data sets.

Applying Dirichlet Process Mixture Models to Compositional Data

Marie Gantz, Research Triangle Institute, viele@ms.uky.edu

Abs: I investigate a dataset providing the composition of trains throughout the United States over a 5 year time span. The goal of the analysis is to cluster the variance routes by the composition of the cargo. To perform the clustering, a dirichlet process mixture model is applied to the compositional data, resulting in several different types of trains being identified. We also provide extensions to ”regression type” compositional settings, where covariates are present.

Functional Clustering of Temporal Microarray Data

Ping Ma, Harvard University, viele@ms.uky.edu

Abs: In this talk, I present a functional clustering method based on a mixture smoothing-spline model. For each cluster, I model its mean profile using a smoothing spline and describe each individuals gene's variation by a parametric random effect. I present an EM algorithm to find the maximum a posteriori. This method automatically takes care of the missing data and infers the number of clusters in the data. I used the developed method to analyze a microarray dataset consisting of the results from 69 individual microarray experiments conducted over the life cycle of the fruit fly. The resulting clusters are validated by examining the statistical over-representation of certain biological functions using the Gene Ontology database. The majority of clusters I obtained are enriched for known and expected biological functions.

DATABASES AND COMPUTERS (Contributed, 4:15-6:00)

Automated Generation of Metadata

Faleh Al-Shameri, George Mason University, falshame@gmu.edu

Abs: The capabilities of generating and collecting data have been increasing rapidly. The computerization of many business and government transactions, and the advances in data collection tools have provided us with huge amounts of data. Millions of databases have been used in business management, government administration, scientific and engineering data management, and many other applications. Massive data sets are collected routinely in a variety of settings in astrophysics, particle physics, genetic sequencing, geographical information systems, weather prediction, medical applications, telecommunications, and sensors. This explosive growth in data and databases has been generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in database, data warehouses, or other information repositories. The methodology used in this research depends on the METANET concept. I consider a heterogeneous collection of massive databases, such as remote sensing data, and text data. The methodology is divided into two sections. The first section is automated generation of metadata, and the second one is the query and search of the metadata. In order to help scientists on searching massive databases and find data of interest to them, a good information system should be developed for data ordering purposes. On my research I am designing a web-based prototype to demonstrate the idea. The prototype will allow scientists to make queries against disparate types of databases. Visualization will play role to discover unexpected correlation and causal relationships, and understand structures and patterns in the massive data.

A Database Design for Assessing Student Learning in an Online Course System

Jun Tan, West Virginia University, jtan@stat.wvu.edu
E. James Harner, West Virginia University, jharner@stat.wvu.edu

Abs: Internet technologies provide convenient ways for both teachers and students to access learning materials, to do exercises, to check grades, etc. The teacher and students are no longer required to be physically present during the learning process. Also, the number of students an instructor can teach can be greatly expanded. However, it may be more difficult for the teacher to know what the students know. Furthermore, how does the student get individual feedback from the teacher? Assessment models help solve these problems.

IDEAL (an Intelligent Distributed Environment for Adaptive Learning) is an online course system developed and used by the Department of Statistics, West Virginia University. IDEAL has the attributes of a typical online course system, such as WebCT. IDEAL has been rewritten from the ground up to enhance all functions. Most importantly, we have redesigned the database to allow statistical assessment models to be implemented. Course content is organized into a hierarchy of cognitive attributes. Exercise and quiz questions are created to target one or more cognitive attributes by forming a tree-based version of the Q-matrix. Cognitively motivated psychometric models, implemented in R, use the database of student scores to estimate, and periodically update the estimates of, cognitive mastery.

Assessing Data Interoperability In Federated Distributed Databases

Daniel N. Owunwanne, Howard University, dowunwanne@Howard.edu

Abs: The introduction of computer networks allowed interconnectivity among data systems which maintain information in their databases. This type of interaction between computer systems was an early and primitive form of information exchange rather than information sharing. Over the years, interconnectivity provided necessary foundations for interoperability.

The ability of a collection of interconnected databases (federated distributed databases) to share and exchange information is termed Database Interoperability. The reason for such form of sharing may be to expand a local database with related information maintained in remote databases as well as to query the remote information units, which may be complementary to, or overlapping with the local information units.

In this paper, two different databases were created in different computing machines and are located in different places. The two databases are Oracle and PostgreSQL. An interface was developed using Java Database Connectivity (JDBC) to connect them. The goal was to establish and demonstrate data interoperability between them. Assessment performance test was established to determine the throughput and response time of the interoperability system discussed in this paper. Recall and precision concepts were also used to know the relevance of the results of the information retrieved during the testing of the system. The system throughput was measured in queries-per-millisecond and was used as principal performance metric. Response time was used as performance indicator. One of the reasons for the performance assessment on this system was to evaluate query generation processes for improvement.

Developing Statistical COM Servers

David R. Lemmon, Penn State University, dlemmon@psu.edu
Joseph L. Schafer, Penn State University, jls@stat.psu.edu

Abs: A highly effective way of disseminating new statistical procedures written in C or Fortran 95 is to package them as COM servers. A COM server is an object-oriented software module which, once it is installed and registered with the Windows operating system, can interact with a wide variety of applications (S-PLUS, R, Excel, SAS, SPSS and MATLAB, to name a few), producing consistent results across these platforms. This presentation will introduce and demonstrate how to create and use COM servers written as modules in Fortran 95.

Clustering Heterogeneously Distributed Data

Eduardo A. Socolovsky, Norfolk State University, e.a.socolovsky@larc.nasa.gov

Abs: The design, implementation and complexity of a hierarchical algorithm to cluster heterogeneously distributed data sets in which the data can have arbitrary dimension is presented. The algorithm allows to jointly cluster independent data sets, treating them as the components of a larger global distributed data set, without transferring all data to a central site. It also permits to adjust to the characteristics of each source data set, and to weight the effect of each data set on the total clustering result according to the relevance or reliability of the source. This is accomplished by introducing a global measure of dissimilarity for distributed data, which is computed as a weighted average of local dissimilarity measures. Each local measure should be tailored to the data at the corresponding site, with the main requirement that it satisfies the triangle inequality. The new sine dissimilarity measure and some of its properties and bounds are introduced, since it can be chosen as a local measure for high- or infinite dimensional data. The global dendrogram for the distributed data set is obtained from the local dendrograms by approximating global measure values from lower and upper bounds for the local measures values. The lower bounds are contained in the local dendrogram nodes, and the critical upper bounds have to be correctly computed by properly connecting the data points and using the triangle inequality. Uniformly using the Euclidean distance for the global and local measures, this approach was pioneered by Johnson and Kargupta (J-K) for heterogeneously distributed data, and Samatova et al. for homogeneously distributed data. Not being able to use different local measures and weighting was a severe constraint of the J-K algorithm for an inmersive mission planning environment application, and it is shown that it can produce faulty upper bounds by following its shortest path connection. A remedy is given, and additionally, it is proven that new unique connections can be built from storing which data points yield the node value in the local dendrograms, and with these new connections substantially better upper bounds are obtained. To construct local dendrograms collecting the required supplementary node information, two very efficient single link clustering algorithms are presented, one with optimal complexity. Pseudo-code detail is given of these algorithms and a recursive algorithm to compute the upper bounds. These results, now entering their implementation and testing phase, stem from an ongoing collaboration between Norfolk State University and NASA LaRC.

APPLICATIONS IN MEDICINE (Contributed, 4:15-6:00)

Improving the Sensitivity of Health Care Cost Predictions Using a Combination of Regression and Classification Procedures

Ogi Asparouhov, MEDai, Inc., OAsparoukhov@MEDai.com

Abs: The prediction of next year overall cost is the most important and frequently used data mining outcome for health insurance companies (health plans). One of the most important applications of these predictions is for identification of high-risk patients (members) with next year cost typically exceeding $10-12,000. These costs could be reduced from 25 to 50 percent, leading to significant savings after implementation of effective care management programs. The sensitivity (for example, the percentage of the top 2 percent of members, according to their predicted cost, that also belong to top 2 percent of members according to their actual cost) is the most appropriate measure for the goodness of these models from a disease management point of view (while R2 is more appropriate from an underwriting point of view). The conventional predictive model involves some regression procedures (linear and nonlinear regressions, neural nets, decision trees etc) that predict a continuous output - next year cost. Of course there are many clusters and transformations of the variables within each cluster. Our research aims to show that involvement of the classification approach (alone or in combination with the regression approach) in the identification of high-risk members leads to an improvement of the overall sensitivity. This experimental study was carried out using MEDais client database incorporating several million members and hundreds of predictors.

Multilevel Classification of Quantitative Cytology Data Using Cumulative Log-Odds Method

José-Miguel Yamal, Rice University, jmy@stat.rice.edu
Dennis Cox, Rice University, dcox@stat.rice.edu

Abs: Multilevel classification is a problem in statistics which has gained increasing importance in many real-world problems but has not yet received the same statistical understanding of the general problem of classification. We are interested in the case where we have measurements at a microscopic level, and yet want to classify at the macroscopic level (e.g. measurements on cells within a patient, yet want to classify the patient). In particular, the case where we have a high-dimensional feature vector on each cell has proved to be a challenging problem. We propose using the Cumulative Log-Odds Method in order to deal with this problem for the detection of cervical neoplasia from quantitative cytology obtained from a pap smear.

Automatic Classification of fMRI Brain Images Using Smooth Asynchrony Maps

Svetlana Shinkareva, University of Illinois at Urbana-Champaign, lana@uiuc.edu
Hernando Ombao, University of Illinois at Urbana-Champaign, ombao@uiuc.edu
Bradley Sutton, University of Illinois at Urbana-Champaign, bsutton@uiuc.edu

Abs: Functional neuroimaging methods are increasingly used in clinical applications to look for indications of disease. We present an explicit discrimination and classification procedure that operates directly on functional imaging time series data to classify a specific subject into one of two groups. These groups could be patient versus control, or for cognitive aging work, younger versus older brains. The presented methodology is a unified feature extraction and classification approach based on temporal synchrony with incorporated spatial information. The method has been successfully validated with simulated data. This methodology could be applied to other modalities that produce a time series of images with some form of spatial resolution.

Modeling Exposures for DNA Methylation Profiles

Kimberly Siegmund, University of Southern California, kims@usc.edu

Abs: DNA methylation is an enzymatic modification of DNA frequently found to be abnormally distributed in cancer. DNA methylation profiles, generated by measuring DNA methylation across a series of CpG regions, are used to classify tumors into disease subtypes. Validation of the classes is typically performed using external information such as exposures or outcomes. We present an extended finite mixture model to estimate the association between exposure and latent disease subtype measured by DNA methylation profiles. Estimates from this model are compared to those obtained from the simpler two-phase approach of first clustering the DNA methylation data followed by associating exposure with disease subtype. When the disease subtypes are distinct leaving low uncertainty in cluster assignment, a two-phase analysis of clustering samples followed by association analysis with novel subgroups gave similar results to the more complex single analysis from an extended mixture model. When there is noise in the cluster analysis, differences in the two approaches emerged. In the naive two-phase approach the estimation of association between exposure and disease subtype is biased toward the null and does not attain the correct coverage level as is typical in problems with measurement error. Estimates from the extended mixture model are unbiased and have the correct coverage probabilities but may require large sample sizes. However, the extended mixture model does not provide an external validation of the clustering result as information on exposure level is used by the model in determining cluster membership of the samples. Using logistic regression to validate groups identified by cluster analysis can give biased estimates of association with invalid estimates of precision when there is uncertainty in cluster assignment. Use of the extended mixture model gives unbiased estimates but is no longer an independent validation of the clustering result.

The Shrinkage Variance Hotelling T-Squared Test for Genomic Profiling Studies

Grant Izmirlian, National Cancer Institute, izmirlian@nih.gov
Jian-Lun Xu, NCI, jianxu@helix.nih.gov

Abs: Designed gene expression micro-array experiments, consisting of several treatment levels with a number of replicates per level, are analyzed by applying simple tests for group differences at the per gene level. The gene level statistics are sorted and a criterion taking into account multiplicity is applied. A caveat arises in that true signals (genes truly over or under expressed) are “competing” with fairly large type I error signals. False positives near the top of a sorted list can occur when genes having very small fold-change are compensated by small enough variance to yield a large test statistic. One of the first attempts around this caveat was the development of “significance analysis of micro-arrays (SAM)”, which used a modified t-type statistic thresholded against its permutation distribution. The key innovation of the modified t-statistic was the addition of a constant to the per gene standard errors in order to stabilize the coefficient of variation of the resulting test statistic. Since then, several authors have proposed the use of shrinkage variance estimator in conjunction with t-type and more generally, ANOVA type tests at the gene level. Our new approach proposes the use of shrinkage variance Hotelling T-squared statistics in which the per gene sample covariance matrix is replaced by a shrinkage estimate borrowing strength from across all genes. It is demonstrated that the new statistic retains the F-distribution under the null, with added degrees of freedom in the denominator. Advantages of this class of tests are (i) flexibility in that a whole family of hypothesis tests is possible (ii) the gains of the above-mentioned earlier innovations are enjoyed more fully. This talk will summarize our results and present careful simulation study benchmarking the new statistic against another recently proposed statistic.

(Intentional blank page.)

BEST OF SIAM DATA MINING CONFERENCE (Invited, 10:15-12:00)

Dynamic Detection, Visualization and Classification of Defect Structures in Molecular Dynamics Simulations

Srinivasan Parthasarathy, Ohio State University, srini@cse.ohio-state.edu
Sameep Mehta, Ohio State University, mehta@cse.ohio-state.edu
Steve Barr, Ohio State University, barr@pacific.mps.ohio-state.edu
Alex Choy, Ohio State University, tschoy@pacific.mps.ohio-state.edu
Hui Yang, Ohio State University, yanghu@cse.ohio-state.edu
Raghu Machiraju, Ohio State University, raghu@cse.ohio-state.edu
John Wilkins, Ohio State University, wilkins@pacific.mps.ohio-state.edu

Abs: Advances in numerical simulation techniques, computer hardware and data collection techniques have resulted in large scale realistic simulations. However, as simulations become more detailed and realistic, scientists are finding that manual analysis of the large datasets produced by these simulations is a non-trivial task. Handling noise and incorporating domain knowledge in computational techniques are some other key challenges which need to be addressed

In this case study, we present a framework for understanding the evolution of anomalous structures (defects) in data generated from Molecular Dynamics (MD) simulations of Silicon (Si) atom systems. These defects can have a undesirable impact on the electrical and mechanical properties of Silicon. We present an algorithm to find the defect structures in the lattice. These detected defects are then verified using visualization techniques like isosurfacing, slicing and volume rendering. We also propose a two-step dynamic classifier which classifies the defect structures in real time. The classifier is not only capable of handling an incoming stream of data but also capable of adding new defect classes on the fly as the simulation progresses. The proposed algorithms are robust and scalable in the size of the atom systems. Each phase is immune to noise, which is characterized after a study of the simulation data. We also validate the proposed solutions by using a physical model and properties of the lattice system. We demonstrate the efficacy and correctness of our approach on several large datasets.

Topic-Driven Clustering for Document Datasets

Ying Zhao, University of Minnesota, yzhao@cs.umn.edu
George Karypis, University of Minnesota, karypis@cs.umn.edu

Abs: In this paper, we define the problem of topic-driven clustering, which organizes a document collection according to a given set of topics (either from domain experts, or as a requirement satisfying users' needs). We propose three topic-driven schemes that consider the similarity between the document to its topic and the relationship among the documents within the same cluster and from different clusters simultaneously. We present the experimental results of the proposed topic-driven schemes on five datasets. Our experimental results show that the proposed topic-driven schemes are efficient and effective with topic prototypes of different levels of specificity.

Gaussian Processes for Active Data Mining of Spatial Aggregates

Chris Bailey-Kellogg, Dartmouth College, `cbk@cs.dartmouth.edu`
Naren Ramakrishnan, Virginia Tech, `naren@cs.vt.edu`
Satish Tadepalli, Virginia Tech, `stadepal@vt.edu`
Varun N. Pandey, Virginia Tech, `vnpandey@vt.edu`

Abs: Active data mining is becoming prevalent in applications requiring focused sampling of data relevant to a high-level mining objective. It is especially pertinent in scientific and engineering applications where we seek to characterize a configuration space or design space in terms of spatial aggregates, and where data collection can become costly. Examples abound in domains such as aircraft design, wireless system simulation, fluid dynamics, and sensor networks. This paper develops an active mining mechanism, using Gaussian processes, for uncovering spatial aggregates from only a sparse set of targeted samples. Gaussian processes provide a unifying framework for building surrogate models from sparse data, reasoning about the uncertainty of estimation at unsampled points, and formulating objective criteria for closing-the-loop between data collection and data mining. Our mechanism optimizes sample selection using entropy-based functionals defined over spatial aggregates instead of the traditional approach of sampling to minimize estimated variance. We apply this mechanism on a global optimization benchmark comprising a testbank of 2D functions, as well as on data from wireless system simulations. The results reveal that the proposed sampling strategy makes more judicious use of data points by selecting locations that clarify high-level structures in data, rather than choosing points that merely improve quality of function approximation.

Clustering with Model-Level Constraints

David Gondek, IBM Watson Research Center, `dgondek@us.ibm.com`
Shivakumar Vaithyanathan, IBM Almaden Research Center, `shiv@us.ibm.com`
Ashutosh Garg, Google Research, `ashutosh@google.com`

Abs: We describe a systematic approach to uncovering multiple clusterings underlying a dataset. In contrast to previous approaches, the proposed method uses information about structures that are not desired and consequently is particularly useful in an exploratory datamining setting. Specifically, the problem is formulated as constrained model-based clustering where the constraints are placed at a model-level. Two variants of an EM algorithm, for this constrained model, are derived. The performance of both variants is compared against a state-of-the-art information bottleneck algorithm on both synthetic and real datasets.

MODELING ALCOHOL ABUSE (Invited, 10:15-12:00)

Ecology of Alcohol and Alcoholism

Ed Wegman, George Mason University, `ewegman@galaxy.gmu.edu`
Yasmin Said, George Mason University, `ysaid99@hotmail.com`

Abs: Alcohol abuse leads to acute outcomes for both society and individuals. Among these, we identify DWI crashes with fatalities, assault and battery, suicide, murder, sexual assault, domestic violence, and child abuse. Alcohol abusers are embedded in a social network that involves the user, family and friends, producers and distributors of alcohol products, law enforcement, the judiciary, remediation, education, detox and treatment facilities, which are coupled to insurance and managed-care programs. This complex network is reminiscent of more traditional biologic ecology systems, hence the name. The basic idea is to formulate a model of this network with the goal of exploring short- and long-term interventions that reduce the overall probability of acute outcomes. The framework we are pursuing is a dynamic agent-based simulation. The basic model is a stochastic directed-graph model that follows agents (sometimes referred to as actors or individuals) through a 24-hour period. T! he stochastic directed graph has two major features that we are developing. First we engage in what we call scenario development. This involves development of scenarios of typical behaviors throughout a day for nonusers, casual drinkers, alcohol abusers, and alcoholics. Associated with these scenarios, we are developing methods for estimation of transition probabilities from state to state during the day reflecting different behaviors and specific to both ethnic groups and geographic location. The proclivity of different ethic, geographic, and job classes to alcohol use and abuse is to be modeled by a hierarchical Bayes scheme.

Modeling Alcohol Abuse and Consequences

Yasmin Said, George Mason University, `ysaid99@hotmail.com`
Ed Wegman, George Mason University, `ewegman@galaxy.gmu.edu`

Abs: We develop a model framework for alcohol, which will provide an assessment for interventions meant to minimize acute outcomes (intentional and unintentional injuries/death.) Acute outcomes are often caused by consumption of ethanol. Our framework is ecological (individual agents and interactions are represented), stochastic (neither individual behavior nor consequences of interventions are certain) and flexible. Constructing the framework raises issues in the domain science of alcohol, statistics, mathematics, and computer science. We provide a time and space dependent stochastic digraph model of alcohol use and abuse. The model is a social network model, which captures the dynamics of alcohol abuse and in particular its acute outcomes. The intent is to study potential interventions and investigate their effectiveness at reducing the overall prevalence of acute outcomes. Current interventions focus on one outcome at a time rather than simultaneously considering all outcomes. The work involves both sophisticated mathematics (stochastic digraphs, Bayesian networks, copulas) as well as intensive data collection.

Drinking and Public Health: The Role for Simulations and Models

William F. Wieczorek, Buffalo State, ysaid99@hotmail.com

Abs: We provide an overview of the effects of drinking on public health and also identify opportunities to diminish the negative impacts caused by drinking. The global burden of disease model is highlighted as the most appropriate method for assessing the overall well-being of the population. The basic measure of public health is the disability-adjusted life year (DALY), which incorporates both premature mortality and continuing illness/disability into a single measure. The leading risk factors for mortality and DALYs, including alcohol use, are presented. The presentation compares the relative roles of the risk factors across developed and developing nations. The results of these comparisons highlight the immensely deleterious impact of alcohol use at the population level. The recognition of the severity of the effects caused by alcohol use provides an opportunity to move public health in a quantum leap fashion by focusing on altering alcohol use and its associated ecological systems. There are relatively limited options for population-level interventions to ameliorate alcohols impact, although some, such as taxes, are quite cost effective. However, it is costly to conduct large field trials of new or innovative alcohol-related interventions. These circumstances provide an opportunity to move the field of alcohol-related public health interventions forward through the development of accurate simulations of the alcohol ecological system. A simulation model would allow for the pre-testing of current and new practices to identify promising interventions that could then be tested in the field.

TIME SERIES, NEURAL NETWORKS, HIERARCHICAL MODELS (Contributed, 10:15-12:00)

Time Series Prediction Based on ARMA and GRNN Technology

Weimin Li, Donghua University, Shanghai, 108wml@mail.dhu.edu.cn
Jianwei Liu, Donghua University, Shanghai, liujw@mail.dhu.edu.cn
Jiajin Le, Donghua University, Shanghai, lejiajin@dhu.edu.cn
Xiangrong Wang, Shandong University of Science and Technology, xrwang@sohu.com

Abs: This article present the time series prediction through combining the time series model ARMA with generalized regression neural network (GRNN).The relative tests testify that the prediction of the ARMA-GRNN is better than that of the GRNN. These indicate that the innovation series inferred from the ARMA contains the statistic characters of time series, such as stock data. And bring more information about the time series to the ARMA-GRNN model made by the article. Obviously, it gives a better prediction.

Discrimination of Locally Stationary Time Series

Wolfgang Polonik, University of California at Davis, wpolonik@ucdavis.edu
Gabriel Chandler, Connecticut College, gabriel.chandler@conncoll.edu

Abs: A novel approach to discrimination of time series is proposed and studied. While being based on the idea of feature extraction, our method is not distance based, and as such avoids the necessity of aligning the time series.

As a motivating example for our approach we consider the problem of discriminating earthquakes and explosions. An time varying autoregressive processes is used as an underlying model. We then define nonparametric measures of the shape of the variance function, and base discrimination on these concentration measures. Mathematical justification for our proposed methodology is also presented.

Bayesian Neural Networks and Variable Selection

R. Adam Molnar, University of Chicago, molnar@galton.uchicago.edu

Abs: Neural networks suffer from a problem of explanation; it is very difficult to determine which variables are important. Additionally, since the network estimates coefficients for each variable in each node, high dimensionality leads to serious over-fitting problems. To deal with these issues, the proposed method attaches a Bayesian prior to the parameters. The model forces many coefficients to be nearly zero through a mixture model based on a switching variable. If the switch is on, the coefficient takes a typical normal distribution, but when off, the coefficient is shrunk to a very small value. The near zero coefficients reduce over-fitting, while the switches provide evidence of importance. Comparative results will be shown.

Hierarchical Bayesian Models for Frequent Terms in Text

Edoardo M. Airoldi, Carnegie Mellon University, eairoldi@cs.cmu.edu
William W. Cohen, Carnegie Mellon University, wcohen@cs.cmu.edu
Stephen E. Fienberg, Carnegie Mellon University, fienberg@stat.cmu.edu

Abs: Most recent statistical approaches to modeling text implicitly assume that informative words are rare. This assumption is usually appropriate for topical retrieval and classification tasks; however in non-topical classification and soft-clustering problems, such as classification by sentiment or author, informative words can be very frequent, and it becomes necessary to reconsider this assumption.

In this paper we present a comprehensive set of statistical learning tools which treat high-frequency words with higher frequencies of occurrence in a sensible manner.

(1) We introduce two hierarchical Bayesian models for classification and soft-clustering based on the Poisson and Negative-Binomial distributions, which retain the desirable properties of simplicity and analytic tractability of the quantities that are relevant for inference purposes, and at the same time allow us to introduce dependence among multiple occurrences of the same word, thus providing a natural notion of context. Our models make use of novel re-parameterizations of the Poisson and Negative-Binomial distributions that provide more intuitive meaning of the parameters. The re-parameterization also lets one adopt the Dirichlet distribution as a natural non-informative prior, and to carry out the inference

conditionally, based on maximum likelihood estimates of a set of crucial parameters, thus enhancing robustness.

Most importantly, the analytic tractability of our models enables efficient inference mechanisms to be devised for more complex language models, such as latent Dirichlet allocation (Blei, Ng, and Jordan, 2003) or author-topic models (Erosheva, Fienberg, and Lafferty 2004), by simply plugging in these more realistic distributions and then updating the formulas—with some necessary approximations. We show, for example, how it is possible to use simple approximations and obtain a lower bound for variational inference in closed form for a soft-clustering version of our models.

(2) We introduce a novel statistic for selecting features according to their importance, the Delta Squared statistic, which helps us avoid over-fitting by making use of sound assumptions about the particular distribution for the occurrence of words. False Discovery Rate arguments are used to control the overall probability of selecting irrelevant words. Again to enable fast inference, we compute the asymptotic distributions of the Delta Square statistic with different degrees of precision, assuming both Poisson and Negative-Binomial word counts. This allows to quickly compute p-values for Delta Square on its approximate distribution using formulas instead of drawing millions of samples.

(3) We demonstrate that on a wide range of classification tasks, our models perform better than widely used models based on the Multinomial distribution, and unlike models based on the Multinomial, yield reasonable log-odds. We also show that the soft-clustering version of our model is able to extract a richer set of latent patterns and fit the data better than latent Dirichlet allocation.

(4) We introduce an exploratory tool that allows us to measure the frequency/variability content of a set of words and map it onto a continuum, from Binomial to Poisson to Negative-Binomial, based on the value of a certain parameter. This tool helps to explain some of the empirical results, and could be used to improve the design of realistic simulations.

MODEL AVERAGING AND ASSESSMENT (Contributed, 10:15-12:00)

A Modest Improvement Towards Seeing the Trees in a Utopian Forest

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

Abs: In recent work I considered the use of the random forest (RF) algorithm the classification and important feature selection within a SELDI-TOF proteomics study in the setting of a cancer prevention trial. The intent of that study was to take a closer look at the statistical properties of feature detection using an “importance measure” via the monte carlo (MC) simulation of “realistic looking” spectra for responders and non-responders to a candidate cancer prevention drug. In the course of that study several observations were made. First, that the variance estimator provided in the current version of the program is incorrect, being of order 1/B, where B is the number of bootstraps, rather than of order 1/n where n is the sample size. A new variance formula has been derived. A second

observation made, using data simulated under the global null hypothesis, was that the standard error normalized importance measures (using the MC variance) were not at all "t" distributed, but were fat tailed instead. Thus when considering an ordered list of standard error normalized importance measures, obtaining p-values using the nomiminal distribution in the Benjamini-Hochberg procedure at a false discovery rate of 10 percent resulted in an observed false discovery rate of 79 percent. This problem was remedied by referring the statistics to the true null distribution obtained via kernel density estimate applied to data simulated under the global null. In this talk I will apply these results to the analysis of a single dataset. Specifically, I compare the new variance formula with MC simulated data in order to verify its correctness. Next, the per feature variances are estimated and used to normalize the importance measures. Finally using the assumption of equality in distribution of the standard error normalized importance measures under the global null, a kernel density estimate of the true null distribution can be obtained via outer bootstrap of the RF procedure on the original dataset having the class to feature relationship noised via random permutation on each outer bootstrap replicate.

Applying Ensemble Learning to Model Quantitative Structure-Activity Relations of Pharmaceutical Molecules

Christopher Tong, Merck Research Labs, christopher_tong@merck.com
Vladimir Svetnik, Merck Research Labs, vladimir_svetnik@merck.com
Ting Wang, Merck Research Labs, ting_wang@merck.com
Andy Liaw, Merck Research Labs, andy_liaw@merck.com

Abs: Ensemble learning is applied to the problem of Quantitative Structure-Activity Relationship (QSAR) modeling for pharmaceutical molecules. This entails using a quantitative description of a molecular structure to predict its biological activity. The two most prominent tree ensemble learning methods, Random Forest (RF) and Stochastic Gradient Boosting (SGB), are studied on ten classification and regression data sets ranging in size and representing various biological activities and molecular structures. The performance of both ensemble methods is compared with that of other techniques often used in QSAR modeling: SVM, KNN, PLS, and Naive Bayes. Our results indicate that ensemble methods are the "safe methods" to be used with a variety of descriptors and small and large data sets. This means that using one of these methods (SGB or RF) will allow reaching either the best or close to the best performance with no or relatively easy parameter tuning. This and additional features, such as descriptor importance and partial dependence functions, make tree ensembles powerful modeling tools particularly suited for QSAR modeling. An example on the use of these features is provided and discussed in detail.

A Robust Meta-Classification Scheme for Cancer Detection

Gabriela Alexe, Institute for Advanced Study, Princeton, galexe@ias.edu
Gyan Bhanot, IBM Research, TJ Watson Center, gyan@us.ibm.com
Gustavo Stolovitzky, IBM Research, TJ Watson Center, gustavo@us.ibm.com

Lilian Chiang, Institute for Advanced Study, Princeton, lchiang@ias.edu
Jorge Lepre, IBM Research, TJ Watson Center, leprej@us.ibm.edu
Ram Ramaswamy, Institute for Advanced Study, Princeton, rama@ias.edu
Babu Vengataraghavan, Institute for Advanced Study, Princeton, babu@ias.edu

Abs: A major challenge in cancer diagnosis from genomic data is to develop accurate and robust classification models which are independent of the analysis techniques used and are able to combine data from different laboratories. We present a novel meta-classification scheme originally developed for phenotype identification from mass spectrometry data which uses a robust combinatorial biomarker selection procedure and integrates the results of several machine learning tools trained on raw and pattern data. We illustrate and validate our method by applying it to distinguish different non-Hodgkin lymphoma and different breast cancer phenotypes from gene array data.

Comparison of Estimators of Generalization Error

Rory Martin, Millennium Pharmaceuticals, rmartin@mpi.com
Kai Yu, Washington University, kai@wubios.wustl.edu

Abs: An important goal in machine learning is to ensure predictive models are robust and replicable. This requires an estimate of a model's true error rate, defined as the conditional generalization error formed by using test points drawn from the same distribution that generated the training data. There are many choices available depending upon whether we draw observations with or without replacement, or sample once, repeatedly, or not at all, and the pros and cons of each are often unclear.

We conducted a Monte Carlo simulation study to illustrate and compare three different estimators of generalization error: bootstrap median, split sample, and resubstitution, and their relationship to true error. Here, "split sample" refers to a single random partition of the data into one pair of training and test samples, a popular scheme. We used stochastic gradient boosting as a learning algorithm, and considered data from two studies for which the underlying data mechanism was known to be complex: (1) a library of 6000 tri-peptide substrates collected for QSAR analysis of the proteasome, and (2) a cardiovascular study involving 600 subjects receiving antiplatelet treatment for acute coronary syndrome.

For each study separately, we applied a two tier sampling design to estimate distributions for true error, bootstrap error, split sample error, and resubstitution error. Despite the highly different nature of the data, similar behavior was displayed in both studies, with the distribution of the bootstrap estimator being closest to that of the true error.

On Evidence Weighted Mixture Classification

Richard Everson, University of Exeter, R.M.Everson@ex.ac.uk
Trevor Bailey, University of Exeter, T.C.Bailey@exeter.ac.uk
Wojtek Krzanowski, University of Exeter, W.J.Krzanowski@exeter.ac.uk
Derek Partridge, University of Exeter, D.Partridge@exeter.ac.uk

Abs: Classifier averaging is a well established method for avoiding the biases inherent in selecting a single parameter-model choice for classification, and has been shown to lead to enhanced predictive accuracy. Bayesian model averaging provides a principled way of achieving a suitable probabilistic combination of classifiers. However, Markov Chain Monte Carlo (MCMC) methods and the Reversible Jump extension (RJMCMC) only easily permit Bayesian model averaging over parameter values within a single structural family of classification models. Combining classifiers with disparate architectures (for example, k-nearest neighbors, radial basis functions and logistic regressors, etc.) is a less tractable problem, both theoretically and computationally.

In this paper we consider the solution of running parallel RJMC chains (one for each structural family of classifiers), and then systematically recombining samples from all these chains in proportion to the evidences (or marginal likelihoods) of each classifier family. The resulting Evidence-Weighted Mixture (EWM) classifier may be viewed as a Multiclassifier System (MCS) in which the final mix of classifiers is self-selected on a sound probabilistic basis that is not biased by choice of either model, or parameter values within model.

A key element in implementing the EWM strategy is an appropriate and computationally feasible technique for calculation of the evidence associated with each classifier family. Marginal likelihood plays a central role in model selection and assessment in the general Bayesian framework and various methods have been suggested to directly or indirectly estimate model evidence. Conventionally this involves the problem of calculating the integral of the likelihood of the data under the model, over the prior distribution of the model parameters. However, in the EWM classifier our interest is on accurate prediction of future, as yet unobserved, feature vectors. In line with standard sequential Bayesian philosophy, it would therefore seem more appropriate to focus on model evidences measured by integrating likelihood over the posterior distribution of the parameters as derived from the training data. A problem then arises in reuse of the same training data set in both deriving the posterior and also calculating the marginal likelihood using this posterior. A possible way forward is to use a form of cross-validation where the available training data are repeatedly partitioned into an evidence set and a training set. We present a straightforward general computational scheme for calculating the evidence based on such an approach and illustrate the efficacy of this scheme in a simple example.

BIOINFORMATICS (Invited, 2:15-4:00)

Statistical Learning Tools for Analyzing Metabolomic Datasets

Xiaodong Lin, University of Cincinnati, linxd@math.uc.edu

Abs: Metabolomic datasets include the quantitative measurements of small molecules, known as metabolites, in a biological sample. These datasets incur many statistical challenges: the number of samples is less than the number of metabolites; there is missing data and non-normal data; there are high correlations among the metabolites. we investigate the use of robust singular value decomposition, recursive partitioning, random forest, SVM and SVM with feature selection to understand a complex metabolomic dataset. We will present the performance of each technique and the importance of feature selection in analyzing this type of datasets will also be discussed.

Learning Variable Covariances via Gradients

Sayan Mukherjee, ISDS and IGSP, Duke University, sayan@stat.duke.edu
Ding-Xuan Zhou, City University of Hong Kong, mazhou@cityu.edu.hk

Abs: The problem of regression and classification can be summarized as learning a function (input-output mapping) from sparse data. In many applications it is of great interest to understand how the input and output variables covary. We introduce a regularization or shrinkage algorithm that estimates the gradient of the regression or classification function. We prove that a representer theorem holds for this algorithm and provide an approximate algorithm that reduces the sizes of matrices required dramatically. We show in a toy example and a gene expression dataset how the gradient can be used to learn variable covariances and structure. A proof of the convergence of the estimated gradient to the true gradient as a function of the number of samples is given.

A Stepwise Structural Equation Modeling Algorithm to Reconstruct Genetic Networks

Grace Shieh, Institute of Statistical Science Academia Sinica, gshieh@stat.sinica.edu.tw
Ching-Yun Yu, Institute of Statistical Science Academia Sinica
Chung-Ming Chen, Institute of Biomedical Engineering, chung@ntu.edu.tw
Juiling Huang, Institute of Statistical Science Academia Sinica

Abs: A model that describes interactions of observed gene-gene interactions and latent factors (for instance, transcriptional factors and other proteins)-genes interactions is proposed. We introduce a stepwise structural equation modeling (SSEM) with BIC and other model selection criteria to learn an optimal structure from microarray data. Simulation studies show that for both 6-gene and 10-gene interaction networks, the proposed algorithm with BIC achieves 97% and 98% true positive rates, respectively provided that sample size =100 and signal noise ratios (SNR) of data is about 2.0 (high). Results of simulation studies of medium (SNR=1.3) with various sample sizes are illustrated.

CLUSTER VALIDATION (Invited, 2:15-4:00)

Additive Tree Fitting of Individual Differences Through (Heuristic) Iterative Projection

H.-F. Koehn, Univ. of Illinois–Urbana-Champaign, hkoehn@cyrus.psych.uiuc.edu

Abs: A new method for the discrete non-spatial representation of individual differences through fitting additive tree structures to three-way two-mode proximity data is proposed. Following the rationale that different individuals base their judgments on the same family of trees with identical topological structure, individual variation is modeled through different branch lengths. However, distinct from existing implementations, such as INDTREES, neither gradient-based optimization, nor a penalty function to account for violations of the four-point condition by the estimated distances is utilized for minimizing the least squares loss function. Instead, the method presented relies on iterative projection onto closed convex sets defined by the constraints of the four-point condition underlying the additive tree model. First, a "communal" tree is identified through application of iterative projection as a heuristic search strategy for the best structural representation of an aggregated proximity data. Individual tree structures, restricted to the identical topology as the communal tree, are then constructed for the individual data matrices through iterative projection with constraints defined by the communal tree structure. An application to judgments of schematic face stimuli illustrates the new method.

Multiobjective Programming Methods for Applied Cluster Analysis

Michael Brusco, Florida State University, mbrusco@garnet.acns.fsu.edu

Abs: There are at least two ways that a multiobjective programming paradigm can improve an applied cluster analysis. First, for the analysis of a single data set, the identification of a clustering solution that produces excellent index values for multiple criteria increases confidence in the solution. Second, in the case of multiple data sets for the same collection of objects, multiobjective cluster analysis can be used to obtain a single clustering solution that provides good index values for each data set. I discuss possible models and applications within this multiobjective paradigm.

Stability Analysis in K-Means Clustering

Douglas Steinley, University of Missouri–Columbia, steinleyd@missouri.edu

Abs: The problem of local optima in K-means clustering is well-documented. To take advantage of this problem, a stability measure based on repeated random initializations of the K-means algorithm is developed. The stability measure is derived by creating an approximate block-diagonal co-occurrence matrix representing the proportion of times pairs of objects are clustered together. It is shown how cluster membership can be derived and validated from the co-occurrence matrix. Additionally, the co-occurrence matrix is extended to detecting fuzzy partitions, influential data points, and determining the number of clusters.

SELECTED IASC PAPERS (Invited, 2:15-4:00)

Data Reduction Using L_p Criteria

Jonathon Schuler, Logos Technologies, jschuler@logostech.net
James E. Gentle, George Mason University, jgentle@gmu.edu

Abs: An important technique in classification and data analysis generally is reduction of the dimensionality of the data by projection onto a lower dimensional subspace that preserves interesting characteristics of the original data. Principal components analysis, for example, is a projection onto spaces that maximize an L_2 norm of the deviations of the data. Although an L_2 norm, because it corresponds to the variance, is the most obvious measure to use in data reduction, more robust reductions are possible with other L_p norms for $p < 2$. In this paper we describe methods for data reduction under general L_p norms and compare their properties with those of standard principal components data reduction.

Empirical Bayes Thresholding in Gene Expression Analysis

Michael G. Schimek, Medical University of Graz, michael.schimek@meduni-graz.at
Wolfgang Schmidt, Medical University of Graz, wolfgang.c.schmidt@gmx.at

Abs: In microarray experiments we are confronted with the problem of high-dimensionality because of tens of thousands of genes involved and at the same time with small sample sizes. For many research tasks involving classification and prediction it is necessary to pre-select a set of differentially expressed genes. This helps improving the performance of the classifier or predictor.

The identification of expressed genes is statistically as well as computationally demanding. A common approach is multiple testing. For each gene a statistic is calculated that is a function of the data. Apart from the type I error (false positive) and the type II error (false negative) compound error measures need to be calculated. As an alternative we propose an empirical Bayes thresholding (EBT) approach for the estimation of possibly sparse sequences observed with white noise (modest correlation is tolerable). A sparse sequence consists of a relatively small number of informative measurements (in which the signal component is dominating) and a very large number of noisy zero measurements. Gene expression analysis fits into this concept. For that purpose we apply a new method outlined in Johnstone and Silverman (2004).

The promising performance of EBT is demonstrated for cDNA measurements.

Fitting a Cox Survival Model on High Dimensional Data

Hans C. van Houwelingen, LUMC, The Netherlands, jcvanhouwelingen@lumc.nl

Abs: When dealing with $p >> n$ in survival data it is natural to add a penalty term to the partial log-likelihood and find the optimal penalty weight by cross-validation. The first problem is how to cross-validate in the partial log-likelihood. The second problem is to find algorithms for fitting the penalized Cox model that are fast enough to allow cross-validation. We use a suggestion by Verweij and Van Houwelingen (1993) to solve the cross-validation problem and the full likelihood instead of the partial likelihood to solve the second problem.

We show the feasibility of this approach on the well-known breast cancer microarray dataset of van de Vijver et al. (2002). Furthermore, we discuss how cross-validated linear predictors can be used when checking the goodness-of-fit of the model and the confounding with clinical prognostic factors

Differential Co-Expression: A New Concept in Analyzing Microarrays

Dennis Kostka, Max Planck Institute for Molecular Genetics, kostka@molgen.mpg.de
Rainer Spang, Max Planck Institute for Molecular Genetics, spang@molgen.mpg.de

Abs: Standard analysis routines for microarray data aim to discover single genes which are differentially expressed. Our approach is complementary: we aim to detect groups of genes which show differential co–expression patterns between two phenotypically distinct sets of expression profiles.

While differential gene expression aims at changes in first order moments (means) we opt for changes in second order moments (covariances). Therefore, in contrast to differential expression, differential coexpression cannot be analyzed in a gene-wise manner. One needs to take into account all possible subsets of genes. The challenge then is to efficiently screen the astronomically large number of such subsets. We suggest to use an additive model for scoring coexpression, which allows for a efficient search.

We present a computationally fast heuristic as an approximation to the solution of the combinatorial problem above. This allows us to find groups of high scoring genes. The method is then demonstrated in the context of simulations and on expression data from a clinical study. The significance of our findings is tentatively assessed via a permutational procedure.

Software is available at http://compdiag.molgen.mpg.de/software/d_coex.shtml

MODEL-BASED/GRAPH-THEORETIC METHODS (Contributed, 2:15-4:00)

Variable Selection for Model-Based Clustering

Nema Dean, University of Washington, Seattle, nemad@stat.washington.edu
Adrian E. Raftery, University of Washington, Seattle, raftery@stat.washington.edu

Abs: In variable selection for model-based clustering the problem of comparing two nested subsets of variables is recast as a model comparison problem, and addressed using approximate Bayes factors. Greedy and headlong search algorithms are proposed for finding a local optimum in model space. The resulting method selects variables (or features), the number of clusters, and the clustering model simultaneously. We present the results of applying the method to several datasets. In general, removing irrelevant variables often improves performance. Compared to methods based on all the variables, variable selection consistently yields more accurate estimates of the number of groups and lower classification error rates, as well as more parsimonious clustering models and easier visualization of results.

Estimation and Selection of Normal Mixture Models Based on Spacings

Yong Wang, University of Auckland, New Zealand, yongwang@stat.auckland.ac.nz

Abs: The determination of the number of components in a heteroscedastic normal mixture model is a known difficult task, mainly due to the unboundedness of the likelihood function near the boundary of the parameter space. In this paper, we address this problem using the maximum product of spacings estimates, which can not only provide a consistent estimate for fixed model dimensions, but also make comparison and selection much easier among models with different dimensions. It is shown that a conventional model selection criterion such as AIC or BIC that penalizes the likelihood function with model dimensions can be readily modified to incorporate spacings, and safely used for all models in comparison, without the need for further scrutiny over the spuriousness of local maxima, as would have to with maximum likelihood.

Model-Based Clustering Toolbox for MATLAB

Angel R. Martinez, Strayer University, angel.r.martinez@navy.mil
Wendy L. Martinez, Office of Naval Research, martinwe@onr.navy.mil

Abs: Model-based clustering takes a finite mixture density estimation approach to finding groups, as well as estimating the number of groups represented by the data. The methodology includes several parts: agglomerative model-based clustering as a way to find partitions to initialize the Expectation-Maximization (EM) algorithm, the EM algorithm to refine the estimate of the density function, and the Bayesian Information Criterion (BIC) to choose the most appropriate model. In this paper, we present a free MATLAB toolbox that implements model-based clustering. The methodology will be explained as we demonstrate the toolbox using a real data set.

Discovering Backbone Structure in Graphs

Juan Lin, Rutgers University, jklin@stat.rutgers.edu

Abs: Data in the form of large sparse graphs are commonly seen today in the fields of internet traffic analysis, citation link analysis, and social network analysis. We present an intuitive highway-traffic model for analyzing large networks. In contrast to algorithms which simply segment the graph, our analysis finds communities in the graph, and in addition extracts relationships between communities in the form of a backbone structure of the graph. Numerical results are presented demonstrating the analysis.

Graph-Theoretic Scagnostics

Leland Wilkinson, SPSS Inc., leland@spss.com
Anushka Anand, University of Illinois at Chicago, aanand2@uic.edu

Abs: Around 20 years ago, John and Paul Tukey developed an exploratory visualization method called scagnostics. While they mentioned their invention in an IMA visualization

workshop, the specifics of the method were never published. Scagnostics stands for Scatterplot Diagnostics. Given a multivariate dataset, the Tukeys developed a set of measures characterizing the distribution of points in each pairwise scatterplot. These measures included the perimeter and area of the convex hull, the curvature of principal curves fitted to the points, the perimeter of selected level curves from kernel density estimates, and so on. By plotting these measures in a scatterplot matrix, the Tukeys were able to identify unusually shaped point clouds in the $p(p-1)/2$ scatterplots.

We introduce a graph-theoretical approach to this problem. We have derived a set of measures based on proximity graphs computed on 2D scatterplots (the convex hull, the minimum-spanning tree, the alpha shape, and a sequence graph). From these graphs, we compute measures of outliers, shape, trend, density, and coherence. Because our basis graphs are subsets of the Delaunay triangulation (or the simple linear sequence graph), our computations are $O(n \log n)$. We further improve computation time by using hexagon binning before computing the triangulation. And because the measures are graph-theoretic, we require no assumptions based on continuous probability distributions. We demonstrate the use of these methods on sample datasets and propose extensions for data-mining applications.

CLUSTERING AND CLASSIFYING TEXT (Invited, 4:15-6:00)

Estimating Probability Mass Functions from Very Sparse Data

Sanjeev P. Khudanpur, Johns Hopkins University, khudanpur@jhu.edu

Abs: A recurring problem in statistical language modeling and clustering of natural language texts is data sparseness. The distribution of words in a document is usually modeled using a non-parametric probability mass function (pmf), that needs to be estimated from sample text. The dimension of such a pmf (vocabulary size) is often tens of thousands, while a document itself may be just a few thousand words long, leading to severe data sparseness problems. Several smoothing or regularization methods have therefore been studied in language modeling, and will be briefly surveyed in this presentation. The majority of time will be devoted to discussing a new method of density estimation. In essence, we will consider as admissible estimates, all pmfs under which the observed word-counts are more likely than any other set of word-counts possible for the same document length. This fundamental shift, from finding a pmf that makes the observed word-counts as likely as possible (the MLE) to simply any pmf that makes the observed word-counts more likely than other possible word-counts, results in a set of pmfs with many desirable properties. We will discuss these properties, as well as a way of choosing one of the admissible pmfs as an estimate. We will present some ongoing work in the application of this method to statistical language modeling.

Knowledge Acquisition from Text

Dekang Lin, Google, Inc., lindek@cs.ualberta.ca

Abs: Text is arguably the richest repository of human knowledge. Two approaches have commonly been adopted in knowledge acquisition from text. One is to define specific patterns and extract instances matching these patterns in a text collection. This has been used to find relationships between words, such as is-a and part-whole. Another approach is based on indirect associations between words in text, as exemplified by many methods for computing word similarity. I will present extension and generalization of the previous methods and show that seemingly deep linguistic or world knowledge may be acquired with superficial statistics.

An Investigation of Text Mining Techniques for the Analysis of Abstracts

David Marchette, Naval Surface Warfare Center, david.marchette@navy.mil

Abs: We describe some text mining techniques based on term-document mutual information and intersection graphs, and apply them to the problem of the analysis of short documents such as abstracts. We illustrate the ideas on abstracts taken from the 2004 Interface, and show how one might use these text processing methodologies to assign abstracts to sessions in a semi-automated fashion.

MODEL-BASED CLUSTERING AND CLASSIFICATION (Invited, 4:15-6:00)

Model-Based Clustering of High-Dimensional Data

Geoff McLachlan, University of Queensland, gjm@maths.uq.edu.au
Richard Bean, University of Queensland, rbean@maths.uq.edu.au

Abs: In this talk, we consider the clustering of high-dimensional data, where the number of observations n is small relative to the number of feature variables p measured on each observation. A common approach to this problem is to use an hierarchical agglomerative procedure, even when there is no reason that the clusters should belong to a hierarchy such as in the evolution of species. In recent times, model-based clustering has become very popular in the scientific literature, although its application to high-dimensional data is not straightforward. Here we provide a model-based approach that is feasible with high-dimensional data. An advantage of model-based clustering is that it provides a sound mathematical framework for clustering. In particular, it provides a principled statistical approach to the practical questions that arise in applying clustering methods, namely, the question of what metric (distance function) to adopt and the question of how many clusters there are in the data. The results are demonstrated on some microarray gene-expression data sets, where the number of features variables (genes) is in the thousands but the number of tissue samples (observations) is less than a hundred.

Quantitating Differences in Two Multivariate Distributions

Guenther Walther, Stanford University, gwalther@stanford.edu

Abs: An important problem in the analysis of flow cytometry data is to provide a quantitative description of how two samples differ. In particular, it is desired to localize regions where the two distributions differ, and to provide confidence statements for the size of the differences. I will show how this can be done with a permutation approach and address some of the computational and methodological issues that arise.

On Potts Model Clustering and Kernel K-Means

Alejandro Murua, University of Washington, murua@stat.washington.edu

Abs: Clustering high-dimensional data sets is becoming a common problem in emerging fields such as bioinformatics (e.g. finding subtypes of cancer in microarray gene expression data to deliver patient-specific treatment), and text mining (e.g. document clustering for fast and relevant search). Hence reliable clustering methods whose performances do not depend on the dimensionality of the data are needed. In this work, we introduce Potts model clustering as a powerful kernel-based clustering method. We built on the work of Blatt, Wiseman and Domany (1996) who, borrowing from known algorithms in physics, used Potts models as a general tool for data clustering. A great advantage of our Potts model clustering methodology over other kernel-based clustering methods, such as kernel K-means, is that it estimates both the clusters and their number simultaneously.

We argue that the key to the success of kernel-based methods lies in their ability to perform clustering through an implicit non-parametric modeling of the density underlying the data. We also show that a slightly modified version of both Potts model clustering and kernel K-means (a penalized Potts model clustering and a weighted kernel K-means, respectively), solve the same clustering problem. Furthermore we introduce an algorithm, a penalized version of the Wolff algorithm, to uncover the cluster structure suggested by penalized Potts model clustering.

The link between kernel-based methods and non-parametric density estimation allows us to propose several estimates of the kernel bandwidths in order to improve the performance of the algorithms. The Markov Chain Monte Carlo nature of Potts model clustering makes it possible to estimate kernel bandwidths under diverse bandwidth smoothing constraints. As a by-product the bandwidths derived from this procedure may be used to obtain a kernel density estimate of the data.

We applied Potts model clustering to gene expression data, and compared our results to those obtained by Gaussian model based clustering, and a non-parametric dendrogram sharpening method.

Part of this work has been done in collaboration with Larissa Stanberry, and Werner Stuetzle, both at the Department of Statistics, University of Washington, Seattle, USA.

MICROARRAYS (Contributed, 4:15-6:00)

Clustering Methods in Microarrays

Lidia Rejto, Statistics Program, University of Delaware, rejto@udel.edu
Gabor Tusnady, Renyi Mathematical Institute of the HAS, Budapest, tusnady@renyi.hu

Abs: The typical data structure in statistical investigation is a matrix. Data that we have in mind have features in common with a questionnaire. The rows are the questions that the investigator poses concerning an investigated phenomena, and the columns correspond to the answers of the different subjects. In microarray analysis the questions are genes or clones, and the answers are the expression levels of the genes in different cell types or under different conditions.

To cluster microarray measurements we group the elements by rows and columns. Grouping by rows reveals the structure and organization of the basic elements of the cell (the enzymes, the membranes, the energy buffers, etc.), while grouping by columns leads to an understanding of the dynamics of life in the cell.

Working with microarray measurements we developed new clustering methods that offer novel insights into the structure of genes and their dynamics.

In this talk we present three different methods: tree clustering, monotone clustering, and partition clustering. Each has its special power to uncover some hidden property of the data. Tree clustering is a special case of the deconvolution problem where an unknown multidimensional distribution is estimated from a sample corrupted by additive Gaussian noise. Partition clustering is an extension of a stochastic model used in graph theory and it is related to simultaneous clustering of genes and conditions developed in recent years. The idea of monotone clustering is new, thus we describe it in more details.

Microarray Analysis: Is an Ordered Gene List Enough?

Leonard B. Hearne, University of Missouri at Columbia, hearnel@missouri.edu
Eric Antoniou, University of Missouri at Columbia, antonioue@missouri.edu

Abs: An ordered gene list is quite adequate for problems where the study is a binary treatment-control design and the biological samples are as genetically similar as possible. The greater the genetic heterogeneity of the samples the larger the sample size must be to control for the spurious inclusion of differentially expressed genes in the gene list. Unfortunately most scientific enterprises can not afford the resources for large n microarray experiments. Also, many scientific questions are more efficiently addressed in a k-group study design. We present techniques which allow iterative refinement of gene list that utilize statistical methods to remove systematic noise from the data, and provide a metric for differential expression. Clustering methods are then used to help prune the gene list by clustering subjects against known groupings. Furthermore, we have additional information about gene co-regulation from other sources. We need clustering methods that utilize this information intelligently in the context of outbred study populations.

Microarray Gene Selection Using Mantel Correlation with K-means

Bill Shannon, Washington University, wshannon@wustl.edu
Brian Steinmeyer, Washington University, steinmeb@ilya.wustl.edu

Abs: We propose the use of a Mantel correlation, coupled with a k-means clustering strategy in order to select important (signal) genes from microarray data. The k-means algorithm is used to partition the data matrix into signal and noise gene subsets. We hypothesize that distance matrices calculated on signal gene subsets will be highly correlated to the distance matrix calculated over all the signal and noise genes. Similarly, distance matrices calculated on noise gene subsets will be uncorrelated to the distance matrix calculated over all the signal and noise genes. The purpose of our method is to significantly reduce the number of non-important (noise) genes to avoid the pitfalls of the large p small n problem inherent in large microarray data. Results from both simulated and applied microarray data are presented.

Genetic Algorithms for Feature Selection using Mantel Correlation Scoring

John Grefenstette, George Mason University, jgrefens@gmu.edu
Kevin Thompson, George Mason University, kthompso@gmu.edu
Brian Steinmeyer, Washington University, St. Louis, steinmeb@ilya.wustl.edu
William Shannon, Washington University School of Medicine, wshannon@wustl.edu

Abs: 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. The Mantel correlation provides a classifier-independent scoring function for selecting subsets of features. In this talk we describe genetic algorithms (GA) for feature subset selection using the Mantel scoring function. The GA is compared with other feature selection approaches on both artificial data sets and gene expression data.

Gene Selection Using Support Vector Machines with Nonconvex Penalty

Cheolwoo Park, University of Florida, cpark@stat.ufl.edu
Hao Helen Zhang, North Carolina State University, hzhang@stat.ncsu.edu
Jeongyoun Ahn, University of North Carolina, Chapel Hill, jyahn@email.unc.edu
Xiaodong Lin, University of Cincinnati, linxd@math.uc.edu

Abs: With the development of DNA microarray technology, scientists can now measure the expression levels of thousands of genes simultaneously in one single experiment. The fundamental problem of gene selection in cancer study is to identify which groups of genes are differentially expressed in normal and cancerous cells, and it leads to a better understanding of genetic signatures in cancer and the improvement on cancer treatment strategies. Though gene selection and cancer classification are two closely related problems, many existing approaches handle them separately by selecting genes prior to constructing the classifier rule. Our motivation is to provide a unified procedure for simultaneous gene selection and cancer classification and achieve high accuracy in both aspects.

The high dimensional low sample size structure of microarray data demands more flexible and powerful statistical tools for analysis. In this talk we introduce a novel type of regularization in support vector machines to identify important genes for cancer classification. A special nonconvex penalty, called the smoothly clipped absolute deviation penalty, is imposed on the hinge loss function in the SVM. By systematically thresholding small estimates to zeros, the new procedure eliminates redundant genes automatically and yields a compact and accurate classifier.

APPLICATIONS (Contributed, 4:15-6:00)

Using Classification of Professionals to Assess Vocational Guidance Counselors and Their Clients

Olga Mitina, Moscow State University, omitina@yahoo.com
Vera Pchelinova, Moscow State University, omitina@yahoo.com
Leonov Sergey, Moscow State University, omitina@yahoo.com

Abs: The problem of classification can be considered in several aspects: mathematical (development of algorithms of extraction of latent classes on the basis of any external evaluations, observable groupings, etc. Studying of the classified objects themselves on the basis of the received classifications (empirical and (or) theoretical), and studying of classifying subjects (their personal or mental features) using classifying as diagnostics procedure. In the given work last-named case is considered. The system of 53 categories determined 53 classes of different profession (objects, purposes, means of work and so on). More then 200 professions were evaluated. Examinees were specialists in vocational guidance who help youth people to chose their professions with difference experience of work: from beginners (junior psychological students, then newly graduated university students have only professional training experience) till high level professional working not less than 10 years. Also high school students who were clients of vocational guidance offices. We wanted to find if there are any steady latent classes of professions in public consciousness, if there is any significant transformation in classes in process of mastering, whether respondents realized classification of professions which exists as implicit stereotypes in their in individual consciousness giving classification on 53 classes, whether it is possible to speak about forming readiness to choice of profession in a course of vocational guidance using results of classification by young client. Solving these applied psychological problems, authors have met accompanying problems: how to prove, that the suggested set of categories (53) on which it is necessary classify professions is necessary and sufficient from the point of view stability and reliability; what scale rang optimum to use: dichotomy, polychotomy (how many gradations?), classification using fuzzy sets concepts.

Visual Assessment of Simple Association Models

Heike Hofmann, Iowa State University, hofmann@iastate.edu

Abs: Links between statistical models and visual displays are still rare, especially for categorical data in a multivariate framework. However, these links do exist and can be utilized in an exploratory setting. Starting from assessing odds ratios visually, interaction effects of variables can be examined using Doubledecker Plots and Mosaicplots. Association models provide a way to describe trends between ordinal variables. In order to link the theory tightly to the displays, global-local odds ratios are used. Examples from real data sets will highlight the usability and applicability of this approach throughout the paper.

Discriminant Function Analysis in Forensic Authorship Attribution

Carole E. Chaski, Institute for Linguistic Evidence, Inc., cchaski@aol.com

Abs: Forensic authorship attribution arises in many types of legal investigations (e.g., homicide, kidnapping, terroristic threatening, patent infringement, copyright infringement). With the Federal and many State courts requiring forensic techniques to be validated and reliable at a specific error rate, a computational approach to forensic authorship attribution is timely, needed and intellectually interesting. Chaski (2001) demonstrated that many stylometric and intuitive variables were not able to identify authors accurately, while syntactic and punctuation variables were. Meanwhile, Stammatatos et al (2001), Tambouratzis et al (2004) and Baayen et al (2002) have each used discriminant function analysis as a classification procedure for the authorship of texts, focusing on 70-some lexical, syntactic and punctuation variables. Accuracy rates from these studies range from 87-89%. In this talk, I present a set of six variables which have obtained 95% accuracy, using cross-validated discriminant function analysis. Similar results were obtained using logistic regression and two ways of decomposing the textual data.

Employing Priors for Classifying High Risk Prison Inmates

Jong-ho Baek, University of California at Los Angeles, jbaek@ucla.edu

Abs: In California, incarceration in the state prison system is in part organized by security level. The higher the security level, the more restrictive the setting. Upon arrival at a reception center, new inmates are scored within a classification system that is used to determine the appropriate level of security. In this paper, we report on the development and testing of a new inmate classification scoring system. Approximately 20,000 inmates took part in a randomized experiment in which half were assigned to their housing using the existing scoring system and half were assigned to their housing using the new scoring system. Our goal is to use these data to produce an accurate mapping of the most problematic prisoners. There were two key issues: 1) provide a new perspective to analyze highly skewed data using a statistical method. 2) take into consideration inmates cost to the prison system that is a result of misconduct.

Measuring Relationships Between Entities in Free Text

Amanda M. White, Pacific Northwest National Laboratory, amanda.white@pnl.gov
Antonio Sanfilippo, Pacific Northwest National Lab., antonio.sanfilippo@pnl.gov
Christian Posse, Pacific Northwest National Lab., christian.posse@pnl.gov
Ryan E. Hohimer, Pacific Northwest National Lab., ryan.hohimer@pnl.gov

Abs: How can we automate the extraction of information about entities and their relationships with each other from a text document? Information analysts read numerous text documents to gain an understanding of the people and organizations described and their relationships and activities. However, time constraints limit the amount of textual information that can be read and summarized. Automating information extraction allows more textual information to be quickly analyzed.

We accomplished automatic entity and relationship extraction by taking advantage of several existing natural language processing technologies and building upon those tools. Our approach requires extracting both the semantic (specifically named entity) information and the syntactic (grammatical) information from the text. When we then combine the syntactic and semantic information, we can create graphs that describe the entities found and characterize the relationships between them.

(Intentional blank page.)

COMPARING CLASSIFICATION METHODS (Invited, 10:15-12:00)

Classifying the Classifiers

Li Li, George Mason University, `lli1@gmu.edu`
James E. Gentle, George Mason University, `jgentle@gmu.edu`

Abs: Methods of classification generally fall into major three classes, those based on regression, those based on neighborhoods, and those based on recursive partitioning. Within these broad classes, there are many details in which the methods vary. Many techniques for classification use various combination of these basic approaches. Some techniques also use meta methods, such as various voting schemes using input from several basic classification procedures. No single classification method is uniformly best, but it is not clear under what conditions a given method can be expected to perform well. In this study, we evaluate and compare various classification methods using both simulated and real data. Guidance in the selection of effective classification methods under various scenarios is provided.

Comparing Neural Networks and Other Multi-Layer Classification Methods

Jill McCracken, Booz Allen Hamilton, `mccracken_jill@bah.com`
Jeremy Flantzer, Booz Allen Hamilton, `Flantzer_Jeremy@bah.com`

Abs: Neural networks and their variants are often considered as a single monolithic classifier, but in fact there are many variations within the structure of a neural network, such as learning parameter, momentum, layer size, number of layers, and convergence thresholds. This talk will analyze the effects of these variations on the data sets.

Comparing Nonlinear Approaches for Classification

Carlos Alzola, George Mason University, `calzola@cox.net`
Yasmin Said, George Mason University, `ysaid99@hotmail.com`

Abs: The data sets were analyzed using a traditional logistic regression model. Restricted cubic splines were implemented to account for non-linearity of the independent predictors. The use of splines puts this method somewhere between the traditional techniques and the more recent developments in data mining. Logistic regression imposes some structure in the data while the splines allow for the modeling of local variations.

Comparing Ensemble Approaches for Classification

James Shine, US Army Topographic Eng. Cntr. and GMU, `statjim@gmualumni.org`

Abs: Ensemble approaches such as bagging, boosting and random forests have achieved significantly more accuracy with several different classification approaches than using a single instance of the approach. Distinguishing which ensemble approaches work the best

is an ongoing area of study. This talk will compare 3 basic ensemble approaches (bagging, boosting, and random forests) on common data sets and present the results.

QUANTILE REGRESSION (Invited, 10:15-12:00)

Quantile Regression: Beyond the Average Man

Roger Koenker, University of Illinois, rkoenker@uiuc.edu

Abs: Much of applied statistics may be viewed as an elaboration of the linear regression model and the methodology of minimizing sums of squares. These methods are ideally suited to deliver estimates of conditional mean functions, but science – and therefore statistics – is also concerned with the analysis of variability around these conditional mean relationships. Quantile regression complements classical least squares methods by providing a tool to estimate conditional quantile functions. By focusing attention on particular slices of the conditional distribution it is possible to offer new evidence on how covariates influence the response at several distinct points of the conditional distribution. The talk will illustrate the methods with some recent applications in biostatistics.

Quantile Volcano Plots for Identifying Significant Genes in Microarray Data

Xia Li, Washington University School of Medicine, xli@im.wustl.edu
Rob Culverhouse, Washington University School of Medicine, rob@ilya.wustl.edu
Bill Shannon, Washington University School of Medicine, wshannon@wustl.edu

Abs: Quantile Volcano Plots are proposed as a modification to standard Volcano Plots to improve identification of genes from microarray experiments with statistical and biological significance. Standard Volcano Plots declare genes to have significantly different expression between two sample types based on both biological difference (absolute log2 (estimated fold change) greater than some constant threshold) and statistical difference (log10 (P value) greater than some constant threshold). Quantile Volcano Plots improve this method by fitting a quantile regression curve to the null distribution of the standard Volcano Plot data and declaring genes significant based on their relationship to this curve. Since the quantile regression curve adapts to the shape of the data, this method avoids the use of constant thresholds for deciding which genes are differentially expressed. In this paper we describe the algorithm and illustrate its use with pharmacogenomic microarray data.

Quantile Regression for Gene Expression Analysis in GeneChip Arrays

Huixia Wang, University of Illinois at Urbana-Champaign, hwang22@uiuc.edu

Abs: We consider the quantile regression approach to linear models with a random effect, with applications to GeneChip data analysis. In particular, we extend the rank score test

for quantile regression to a class of mixed models. The rank score test can be carried out for hypotheses at a single quantile level or jointly at several quantile levels. The null hypothesis on the mean may also be considered together with the hypotheses on quantiles. The proposed tests are found useful in GeneChip studies for detecting differentially expressed genes and for suggesting candidate genes for further studies.

SPECTRAL METHODS IN DATA ANALYSIS (Invited, 10:15-12:00)

From Text Data Mining to Gene Expression Mining and Back Again

Jeff Solka, Naval Surface Warfare Center, jeffrey.solka@navy.mil
Brandon Higgs, George Mason University, bhiggs@mitre.org
Jeffinfer Weller, George Mason University, jweller@gmu.edu

Abs: This talk will discuss recent work at applying the bipartite bipartition methodology of (Dhillon 2001) to text and gene expression data. The methodology has the nice property that it simultaneously clusters on document (samples) and words (genes). Work will be discussed that extends Dhillon's proposed bipartite methodology to perform a tree-based clustering of gene expression data. The modified methodology was applied to the Golub ALL/AML gene expression dataset. The ability of this technique to reveal disease relevant genes was evaluated using a simple co-occurrence analysis on AML/ALL literature. This completes the loop from text analysis to gene expression analysis and back again to text analysis.

Dissimilarity Matrices and Spectral Projections

Elizabeth Leeds, Naval Surface Warfare Center, leedsem@nswc.navy.mil
David J. Marchette, Naval Surface Warfare Center, david.marchette@navy.mil

Abs: This talk focuses on spectral decomposition methods and the similarities and differences between them. The purpose is to illustrate the effects of utilizing an appropriate dissimilarity measure for a dataset. We begin by examining principal components analysis on a synthetic dataset and outlining how this is equivalent to multidimensional scaling on a Euclidean distance matrix created from the data. Next we look at dissimilarity matrices created from measures other than Euclidean distance such as the Laplacian of a graph induced by the data and the proximity of the data as viewed by a random forest. We examine these dissimilarities for several types of data in order to obtain a better idea of the effect of the dissimilarity measure on the outcome of the spectral decomposition.

Co-clustering of Social Network Data

John Rigsby, Naval Surface Warfare Center, john.rigsby@navy.mil
Jeffrey Solka, Naval Surface Warfare Center, jeffrey.solka@navy.mil

Abs: Social Network Analysis is the study and analysis of groups as social entities to attempt to mathematically understand the interactions of people and groups in society. This analysis includes concepts such as nodes and ties, groups, subgroups, cliques, social prominence and rank, centrality, and prestige. Often data dealing with multiple modalities is clustered one at a time. This paper covers a spectral graph method for co- clustering multiple modes at the same time. I will use the standard social network data set, Davis, Gardner, and Gardner social events attendance, to demonstrate co-clustering methods. Co-clustering is very useful not only because it turns a two step process into a one step process, but it also shows you the multi-mode relationships between different sets of actors.

MULTIDIMENSIONAL SCALING AND ET ALIA (Contributed, 10:15-12:00)

Multidimensional Scaling Algorithms for Large Data Sets

Michael W. Trosset, College of William & Mary, trosset@math.wm.edu
Patrick J.F. Groenen, Erasmus University Rotterdam, groenen@few.eur.nl

Abs: We consider the problem of embedding dissimilarity information in a low-dimensional Euclidean space when N, the number of objects to be embedded, is large. This problem challenges the computational capabilities of traditional multidimensional scaling algorithms. Our approach uses $O(N)$ operations to construct a plausible initial configuration, then uses a fixed number of iterations of a new diagonal majorization algorithm to decrease the raw stress criterion. By using just $O(N)$ of the $N(N-1)/2$ dissimilarities, we obtain reasonable embeddings in just $O(N)$ operations.

Local Multidimensional Scaling for Nonlinear Dimension Reduction

Lisha Chen, The Wharton School, University of Pennsylvania, lisha@wharton.upenn.edu
Andreas Buja, The Wharton School, University of Pennsylvania, buja@wharton.upenn.edu

Abs: In recent years, there has been a marked resurgence of interest in nonlinear dimension reduction methods. Local Linear Embedding and Isomap are among the most successful ones to recover meaningful low-dimensional structures hidden in high-dimensional data. Both of them utilize local neighborhood information to construct a global embedding of the manifold.

In this talk, we will introduce a new nonlinear dimension reduction method, namely Local Multidimensional Scaling. Traditional MDS, as PCA, is designed to recover the submanifold which is embedded linearly, or almost linearly. Local MDS only uses local information from user-chosen neighborhood and achieves its optimal configuration by properly weighting a repulsive force between non-neighboring points. A newly defined family of stress functions provides the users with great flexibility in achieving desirable configurations. Our method does a good job in spreading out the famous Swiss roll and in some of the other illustrative datasets used in the LLE and Isomap papers. The connection between Local MDS and some graph drawing methods will also be discussed.

Maximal Data Piling in Discrimination

Jeongyoun Ahn, University of North Carolina–Chapel Hill, jyahn@email.unc.edu
J. S. Marron, University of North Carolina–Chapel Hill, marron@email.unc.edu

Abs: In a binary discrimination problem, a linear classifier finds a linear hyperplane that separates two classes by partitioning the data space. Especially in a High Di-mension Low Sample Size (HDLSS) setting, there are linear separating hyperplanes such that the projections of the training data points onto their normal direction vectors are identically zero, or some non-zero constant. Of interest in this paper is a linear separating hyperplane such that the projections of the training data points from each class onto its normal direction vector have two distinct values, one for each class. This direction vector is uniquely defined in the subspace generated by the data. A simple formula is given to find this direction. In non-HDLSS settings, this direction vector is the same as the Fisher Linear Discrimination direction vector.

Estimating the Sparse Directions in the Effective Dimension Reduction Space

Zhihua Qiao, The Wharton School, Univ. of Pennsylvania, zhqiao@wharton.upenn.edu
Jianhua Huang, Texas A&M University, jianhua@stat.tamu.edu

Abs: Under the regression model $y = f(\boldsymbol{b}'\boldsymbol{x}, e)$ where $\boldsymbol{x}$ is a p-dimensional vector and $\boldsymbol{b}$ is a $p \times k$ matrix, the aim of dimension reduction is to find the space spanned by the column vectors of b, called effective dimension reduction space (edr). Here the reduced predictor $\boldsymbol{b}'\boldsymbol{x}$ has a lower dimension k. SIR (Li 1991), SAVE (Cook 1991), PHD (Li 1992), and a new class of methods in Ye and Weiss (2003) estimate the edr space by constructing a matrix whose eigenvectors are in the edr space. In this paper we propose an estimate of the edr space when the true column vectors of $\boldsymbol{b}$ are sparse (having lots of zero components). A modification of sparse PCA algorithm is used to implement our method. The sparse estimate significantly improves the result when there is high correlation between the explanatory variables, in which case the original methods can yield unstable estimates.

Uncovering Curvature in Data

Jesse Spencer-Smith, University of Illinois, jbspence@uiuc.edu

Abs: Most applications of multidimensional scaling (MDS) assume that data can be modeled within a flat geometry (Euclidean or city-block spaces). MDS provides a good fit for many data sets. This would seem to imply that many of the spaces of interest are, indeed, flat. But does a good MDS fit (low stress, proper dimensionality) logically imply the spaces are flat? The present work demonstrates that this is not the case, and that MDS flattens

curvature that might be present in data while still showing low stress, returning a distorted representation. Ill describe new tools to uncover curvature in data, and discuss a classic data set that demonstrates curvature.

BEST OF THE JOURNAL OF CLASSIFICATION (Invited, 2:15-4:00)

Looking at Different Options Within the COSA Clustering Algorithm

Jacqueline J. Meulman, Leiden University, meulman@fsw.leidenuniv.nl
Jerome H. Friedman, Stanford University, jhf@stanford.edu

Abs: COSA stands for "clustering objects on subsets of attributes," and its motivation was given by consideration of data of a very special kind, i.e., data sets where the number of attributes (in the columns) is much larger than the number of objects (in the rows). One application is in genomics, where we deal with gene expression data in micro arrays, with very many genes (say, 1,500-40,000), and very few objects (say, 20-250). The research area of genomics (study of genes and their functions) has been subsumed with proteomics (study of proteins and their functions), and metabolomics (study of metabolic profiles of cells, etc.) under the name systems biology (e.g., see Van der Greef et al., 2003). When we have a large numbers of attributes, objects are very unlikely to cluster on all, or even a large number of them. Indeed, objects might cluster on some attributes, and be far apart on all others, and our task is to find that (unique) set of attributes that a particular group of objects is clustering on. Common data analysis approaches in systems biology are to cluster the attributes first, and only after having reduced the original many-attribute data set to a much smaller one, one tries to cluster the objects. The problem here is that we would like to select those attributes that discriminate most among the objects (so we have to do this while regarding all attributes multivariately), and it is usually not good enough to inspect each attribute univariately, because important clustering attributes usually don't have their influence as single agents, but in a (small) group. Therefore, two tasks has to be carried out simultaneously: cluster the objects into homogeneous groups, while selecting different subsets of variables (one for each group of objects). The attribute subset for any discovered group may be completely, partially or nonoverlapping with those for other groups. To limit the search space, and to focus on particular aspects of the data, COSA (Friedman and Meulman (2004b) incorporates targeting. For example, we might seek clusters of samples that have preferentially high or low expression values on subsets of genes. To avoid local optima, it is shown in Friedman and Meulman (2004a) that we need to start with the inverse exponential mean (rather than the arithmetic mean) of the separate attribute distances. By using a homotopy strategy, the algorithm creates a smooth transition of the inverse exponential distance to the mean of the ordinary Euclidean distances over attributes. In Friedman and Meulman (2004a) it is asserted that the homotopy approach is important for the theoretical results, but does not matter much in practice. The current paper will study the use of the homotopy parameter in more detail.

A Hierarchical Methodology for Class Detection Problems with Skewed Priors

Christopher K. Eveland, Equinox Corporation, eveland@equinoxsensors.com
Diego A. Socolinsky, Equinox Corporation, diego@equinoxsensors.com
Carey E. Priebe, Johns Hopkins University, cep@jhu.edu
David J. Marchette, Naval Surface Warfare Center, david.marchette@navy.mil

Abs: We describe a novel extension to the Class-Cover-Catch-Digraph (CCCD) classifier, specifically tuned to detection problems. These are two-class classification problems where the natural priors on the classes are skewed by several orders of magnitude. The emphasis of the proposed techniques is in computationally efficient classification for real-time applications. Our principal contribution consists of two boosted classifiers built upon the CCCD structure, one in the form of a sequential decision process and the other in the form of a tree. Both of these classifiers achieve performances comparable to that of the original CCCD classifiers, but at drastically reduced computational expense. An analysis of classification performance and computational cost is performed using data from a face detection application. Comparisons are provided with Support Vector Machines (SVM) and reduced SVMs. These comparisons show that while some SVMs may achieve higher classification performance, their computational burden can be so high as to make them unusable in real-time applications. On the other hand, the proposed classifiers combine high detection performance with extremely fast classification.

Estimating the Cluster Tree of a Density

Werner Stuetzle, University of Washington, wxs@stat.washington.edu

Abs: Clustering problems occur in may domains, from genomics and astronomy to document analysis and marketing. The general goal is to identify distinct groups in a collection of objects. To cast clustering as a statistical problem we regard the feature vectors characterizing the objects as a sample from some unknown probability density. The premise of nonparametric clustering is that groups correspond to modes of this density. Building on ideas of David Wishart and John Hartigan I will introduce the cluster tree of a density as a summary statistic reflecting the group structure, and I will describe methods for estimating the cluster tree.

NON-NUMERIC DATA ANALYSIS (Invited, 2:15-4:00)

Conditional Independence Modeling for Categorical Anomaly Detection

Chad Scherrer, Pacific Northwest National Laboratory, Chad.Scherrer@pnl.gov
Nathaniel Beagley, Pacific Northwest National Laboratory, Nathaniel.Beagley@pnl.gov

Abs: The concept of stochastic conditional independence is central to a variety of modeling methods such as Bayesian networks, Markov random fields, and hidden Markov models. These types of models have become increasingly popular in recent years, both in statistical and computer science literature, in concert with the dramatic improvement of computational resources for collection and analysis of multivariate data. We will present an alternative method that generalizes all of these approaches, followed by a discussion of the applicability to anomaly detection.

Usage-Based Evolution of Visual Analysis Tools

Elizabeth Hetzler, Pacific Northwest National Lab, `beth.hetzler@pnl.gov`
Stuart Rose, Pacific Northwest National Lab, `stuart.rose@pnl.gov`
Dennis McQuerry, Pacific Northwest National Lab, `mcq@pnl.gov`
Pat Medvick, Pacific Northwest National Lab, `patricia.medvick@pnl.gov`

Abs: Visual analysis tools have been developed to help people in many different domains effectively explore, understand, and make decisions from their information. Challenges in making a successful tool include suitability within a user's work processes, and tradeoffs between analytic power and tool complexity, both of which impact ease of learning. This paper describes our experience working with users to help them apply visual analysis tools in several different domains, and examples of how the tools evolved significantly to better match users' goals and processes.

Scenario Analysis for Homeland Security

Olga Anna Kuchar, Pacific Northwest National Lab., `Olga.Kuchar@pnl.gov`
George Chin Jr., Pacific Northwest National Lab., `George.Chin@pnl.gov`
Paul Whitney, Pacific Northwest National Lab., `paul.whitney@pnl.gov`
Katherine Johnson, Pacific Northwest National Lab., `Katherine.Johnson@pnl.gov`
Mary Powers, Pacific Northwest National Lab., `Mary.Powers@pnl.gov`

Abs: One of the more popular techniques used by intelligence analysts for organizing their knowledge about a particular hypothesis or scenario is the production of a link chart, a nodes-and-edges graph in which the nodes tend to be entities, and the edges indicate salient relationships. Analysts then decipher and identify patterns and relationships among the information to formulate a decision. The process of intelligence analysis contains a cognitive process and a collection of techniques. This process seems to be individualistic, but our team has concluded that cognitive similarities can be extracted and abstracted to provide a new tool to support an analysts cognitive process with link charts. In our presentation, we propose the Theory of Intelligence Network Analysis and its relevance to the intelligence process. Based on our interviews with intelligence analysts, we extracted key features that they use in understanding link charts. We then extended this theory to graph properties and created a methodology of finding similar scenarios to aid intelligence analysts. An understanding of this theory and its practical implications will be presented and demonstrated using our scenario analysis tool.

DISCRIMINATION AND DATA MINING (Contributed, 2:15-4:00)

Choosing Weights for Nearest-Neighbor Classification with Linear Programming

Samuel E. Buttrey, Naval Postgraduate School, buttrey@nps.edu
W. Matthew Carlyle, Naval Postgraduate School, mcarlyle@nps.edu

Abs: Nearest-neighbor classification is a widely-used and successful technique. But it is difficult to know how to weight variables to take into account both their inherent scale (units of measure, say) and their relative importance. We describe a scheme in which we choose weights that are optimal (in the sense of minimizing misclassification rate) for nearest-neighbor classification by casting the problem as a mixed-integer linear program. Since the complete problem grows in size very quickly as the number of observations increases, we also describe a heuristic approach that uses a sequence of linear programs to produce good solutions in reasonable time. Examples from real-life data sets are given.

Classification Via Interpoint Distance Profiles

Jayson D. Wilbur, Worcester Polytechnic Institute, jwilbur@wpi.edu

Abs: In general, nearest-neighbor methods classify objects based on their distance from training observations in each group. These rules are flexible because they do not assume any particular distributional form for the data. However, they generally perform poorly for high-dimensional data. In this talk a method for classification via interpoint distance profiles is presented which maintains the flexibility of nearest neighbor methods, but performs well in high-dimensions. 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 genotypic and phenotypic profiles. Applications of the proposed methodology to source tracking data will be presented as time permits.

Polya Tree Priors for Classification Error Distributions

Andrew Neath, SIU Edwardsville, aneath@siue.edu

Abs: The classification problem is characterized by the need to predict a response category based on measurements of one or more input variables. A typical example is the medical diagnosis setting where the presence or absence of disease is inferred from the results of a diagnostic test. The classification rule is applied to a forecasting sample of observed inputs with responses to be predicted. Information on prevalence and error rate probabilities is available, but values are not known with certainty. Thus, one can continue to learn from the forecasting sample. Polya tree distributions provide a natural model for updating the information on misclassification errors.

SIP Load Test Automation based on Data Mining Algorithms

Arta Doci, UCDHSC, Arta.Doci@cudenver.edu

Abs: Session Initiation Protocol (SIP) is declared as the protocol of choice on voice over IP. The protocol has become very popular due to its simplicity and range of applications that it provides. One of SIP applications is instant messaging, which can handle many simultaneous users. This paper describes a tool that automatically analyses the call/chat/user volumes generated on busy hours in SIP instant messaging. The tool is written in Perl and it performs its analysis through data mining algorithms. The tool provides reports on performance metrics and makes suggestions on how to improve them.

CLUSTERING METHODS AND APPLICATIONS (Contributed, 2:15-4:00)

Imputation-Free Robust Clustering Using Soft Constraints

Nan Lin, Washington University in St. Louis, nlin@math.wustl.edu

Abs: Clustering is one of the most extensively used methods on analyzing multivariate data. Most existing clustering methods can not be applied to data sets with missing values. Common solutions are either to discard the observations with missing values or to impute the missing values before clustering. However, imputations often require assumptions that may not be true in practice. Another practical issue of a clustering method is robustness. For data sets that contain outliers, robust clustering methods are needed to produce reliable results. We developed a clustering algorithm that does not required data imputation and is also robust to outliers. This method is a modification of the data depth-based robust clustering algorithm (Jornsten 2004) using the soft constraint method (Wagstaff 2004). This method clusters subjects by optimizing an objective function that consists of two components. One component is a robust criterion function based on the fully-observed features and the other component is a penalty term given by the soft constraints based on the features with missing values. The advantages of this method are illustrated by both simulation studies and real data examples.

Applications of Parametric and Nonparametric Bayesian Predictive Clustering

Fernando A. Quintana, Pontificia Universidad Católica de Chile, quintana@mat.puc.cl

Abs: We consider probability models for partitions of a set of n elements using a predictive approach, i.e., models that are specified in terms of the conditional probability of either joining an already existing cluster or forming a new one. The inherent structure is motivated by hierarchical models of either parametric or nonparametric nature. Comparison of partition structures is discussed. Cluster construction is considered in the context of several applications, including comparison of discrete distributions and density estimation.

Propensity Scoring and the LATE Distribution from Unsupervised Clustering

Nicholas Lewin-Koh, Eli Lilly and Company, nikko@lilly.com
Obenchain, Robert, Eli Lilly and Company, ochain@lilly.com

Abs: In nonrandomized, observational studies, treatment choices are usually unbalanced relative to measured baseline characteristics of the patients. Statisticians have traditionally tried to compensate for biases using numerical estimates of propensity scores from discrete choice models, such as logistic regression. We explore a new form of patient matching based upon hierarchical clustering of patients. We explore the use of numerous, relatively small clusters with strong near neighbor properties on an appropriate metric. Such clusters usually reflect population heterogeneity as well as exhibit considerable small-sample variation. In our talk, we will first address appropriate metrics when patient characteristics are mixed continuous and categorical variables. We then argue that the resulting distribution of local treatment effects contains key information about possible differential response to treatment.

Profiling Price Dynamics in Online Auctions Using Curve Clustering

Wolfgang Jank, University of Maryland, wjank@rhsmith.umd.edu
Galit Shmueli, University of Maryland, gshmueli@rhsmith.umd.edu

Abs: Electronic commerce, and in particular online auctions, have received an extreme surge of popularity in recent years. While auction theory has been studied for a long time from a game-theory perspective, the electronic implementation of the auction mechanism poses new and challenging research questions. In this work, we focus on the price formation process and its dynamics. We present a new source of rich auction data and introduce an innovative way of modelling and analyzing price dynamics. We represent auctions as functional objects by accommodating the special structure of bidding data. We then use functional data analysis to characterize different types of auctions. Our findings suggest that there are several types of dynamics even for auctions of comparable items. By profiling these differences with respect to features associated with the auction format, the seller and the winner we find new relationships between dynamics and auction settings, and we tie these findings to the existing literature on online auctions. Time permitting, we also present Auction Explorer Suite, an interactive visualization tool for exploring patterns in online auction dynamics.

Clustering for Measurement Error in Expenditure Survey Data

John Dixon, Bureau of Labor Statistics, dixon_j@bls.gov

Abs: Expenditure data from the Consumer Expenditure Surveys is used to represent American consumption in a number of economic indicies. The data is thought to suffer from recall error. Cluster analysis is used to try to identify patterns of response which may be attributed to similar patterns of error.

COMPUTATIONAL BIOLOGY (Invited, 4:15-6:00)

Spike and Slab Gene Selection for Multigroup Microarray Data

J. Sunil Rao, Case Western Reserve University, sunil@nalini.EPBI.CWRU.edu

Abs: DNA microarray data presents high-throughput information about a cell's proteomic composition (using nuclear RNA) and thus biologic insight into molecular differences between cells. Analysis of microarray data is challenging not only because large amounts of information are collected using relatively small sample sizes, but also because the data can be non-standard. A proper analysis must take this into account and also be statistically rigorous. In this talk I present new theory for rescaled spike and slab models and discuss their application to multigroup microarray data. The methodology will be illustrated on a large microarray repository of samples from different clinical stages of metastatic colon cancer. This is joint work with Hemant Ishwaran of the Cleveland Clinic Foundation.

Bayesian Infinite Mixture Model-Based Clustering of Functional Genomics Data

Mario Medvedovic, University of Cincinnati, medvedm@UCMAIL.UC.EDU

Abs: Unsupervised identification of patterns in microarray data has been a productive approach to uncovering relationships between genes and the biological process in which they are involved. Infinite Bayesian mixtures based clustering is an alternative to the heuristic or finite mixture model clustering methods that are commonly used in this context. This approach shares the benefits of the finite-mixture modeling in pooling information from the whole dataset in the process of assigning genes to clusters. On the other hand, it avoids the pitfalls of the process of identifying the "correct" number of mixture components by averaging over models with all possible number of clusters. Clusters of co-expressed genes are formed based on the posterior distribution of clusterings generated by a Gibbs sampler. This distribution can also be used to assess the statistical confidence in identified patterns. In this talk, the infinite mixture model is described and the examples of cluster analyses in which this approach provided a critical improvement in the precision are discussed.

Classification and Regression-Based Approaches to Protein Structure Prediction

Jarek Meller, Cincinnati Children's Hospital Research Foundation, jmeller@chmcc.org
Rafal Adamczak, Cincinnati Children's Hospital Research Foundation, adaq83@chmcc.org
Michael Wagner, Cincinnati Children's Hospital Research Foundation, mwagner@chmcc.org
Aleksey Porollo, Cincinnati Children's Hospital Research Foundation, aporollo@chmcc.org

Abs: Computational protocols for protein structure and function prediction play an essential role in post-genomic efforts to transform the available genomic data into a rich source of medically relevant hypotheses. One example is the prediction of the level of solvent ex-

posure of amino acid residues in proteins, which enables enhanced overall protein structure prediction, classifications of mutations and polymorphisms and identification of functionally important sites. In this talk we discuss accurate novel regression-based approaches to the problem of predicting real valued relative solvent accessibility. Two classes of methods, namely Support Vector Regression and Neural Network-based regression models are compared. We also discuss further applications to improved prediction of protein-protein interaction sites and membrane domains.

GRAPH-THEORETIC PATTERN RECOGNITION (Invited, 4:15-6:00)

Juggling: Ensembles of Class Cover Classifiers

Jason DeVinney, Center For Computing Sciences, jgdevin@super.org
David Marchette, Naval Surface Warfare Center, david.marchette@navy.mil
Carey Priebe, Johns Hopkins University, cep@jhu.edu

Abs: The class cover problem (CCP) is a multi-class set covering problem where the covering sets are constrained to be balls under some dissimilarity measure. Given a collection of data from more than one class, the goal of the CCP for some target class is to find a small set of balls whose union contains all (or most) of the target class points and contains none (or few) of non-target class points. We explore a new ensemble variant of CCP classification.

Geometry of Learning: from Graphs to Continuous Spaces

Mikhail Belkin, University of Chicago, misha@cs.uchicago.edu

Abs: I will discuss the geometric framework for pattern recognition based on the Laplacian operator associated to a point cloud, which connects graph-based methods and geometry of continuous probability distributions. This framework has been applied to data representation, clustering, regression and classification. One of the most promising applications is using unlabeled data to improve classification/regression performance.

Local Intrinsic Dimension Estimation with kNNGs

Alfred Hero, University of Michigan, hero@umich.edu
Jose Costa, University of Michigan, jocsta@eecs.umich.edu

Abs: Much high dimensional data evolves on lower dimensional structures. The specification of local intrinsic dimension allows these structures to be characterized in a spatially dependent manner, e.g. revealing changes in intrinsic dimension over parts of data space. We will present methods of local intrinsic dimension estimation that use growth rates of kNN graphs as discriminants. These estimates have provable convergence properties and can be applied to many applications where data is supported on structures of mixed dimensions.

ALGORITHMS (Contributed, 4:15-6:00)

Parallel Computation of the kth Nearest Neighbor Estimate of the Entropy of Molecules Using Circular Distances

E. James Harner, West Virginia University, jharner@stat.wvu.edu
Jun Tan, West Virginia University, jtan@stat.wvu.edu
Shengqiao Li, West Virginia University & NIOSH, shli@stat.wvu.edu

Abs: 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.

Last year we reported (2004 Interface) our implementation of two parallel algorithms for computing the kth nearest neighbor distances to estimate entropy for very large data sets obtained from molecular dynamics (MD) simulations. On each processor, we use the ANN method (Arya et al., Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, 1994, 573-582) for computing kth nearest neighbor Euclidean distances. However, for circular probability models, it is more natural and more efficient to use circular distances. We have modified the source code of ANN to incorporate two circular distance measures. So far we have addressed the low-dimensional cases in which the estimates of entropy can be found in closed form.

Characterizing the Solution Path of Multicategory Support Vector Machines

Zhenhuan Cui, Ohio State University, zhenhuan@stat.ohio-state.edu
Yoonkyung Lee, Ohio State University, yklee@stat.ohio-state.edu

Abs: The algorithm of fitting the entire regularization path of the support vector machine (SVM) was recently proposed by Hastie et al. (2004). It allows effective computation of solutions and greatly facilitates the choice of the regularization parameter that balances a trade-off between complexity of a solution and its fit to data. Extending the idea to more general setting of the multiclass case, we characterize the coefficient path of the multicategory SVM via the complementary conditions for optimality. The extended algorithm provides a computational shortcut to attain the entire spectrum of solutions from the most regularized to the completely over-fitted ones.

Accelerating Linearly-Convergent Algorithms

Tim Hesterberg, Insightful Corp., timh@insightful.com

Abs: EM and other linearly-convergent algorithms can be very slow to converge. This can be sped up by Aitken acceleration, a step-lengthening method in which the direction between successive parameter vectors is chosen by vanilla EM, but the step size is modified. Aitken is particularly effective when convergence is dominated by a single large eigenvalue, with other eigenvalues near zero. For other situations there are multivariate versions of Aitken, but they are more complicated and tricky. We propose a multiple-univariate version of Aitken acceleration, in which a sequence of step length factors is used to speed convergence for all eigenvalues, without explicitly identifying the eigenvalues.

Alternative Visualization of Andrews' Curves

Wendy L. Martinez, Office of Naval Research, martinwe@onr.navy.mil
Angel R. Martinez, Strayer University, angel.r.martinez@navy.mil

Abs: Andrews' curves are a means of transforming each observation into a continuous function. The transformation can be any of a set of orthogonal functions, with the most common one using sines and cosines. In this paper, we first describe Andrews' curves, provide examples using real data sets, and discuss some of the limitations of visually exploring the data using this method. Then we present alternative visualizations for Andrews' curves that address these problems.

Optimal Linear Combination of Longitudinal Markers for Disease Classification

Ming Ji, San Diego State University, mji@mail.sdsu.edu

Abs: Optimal linear combination of multiple markers is a convenient method for combining multivariate correlated markers to classify subjects into a dichotomous disease outcome. The optimal linear combination applies for very general continuous multivariate distributions and shows that combining more multiple markers always leads to a larger area under the ROC curve. Disease classification based on longitudinal markers is an important clinical application. Some examples include: predicting AID conversation using longitudinal CD4 and viral load data; predicting prostate cancer based on longitudinal PSA measures and predicting Alzheimers disease based on multiple neuropsychological test scores. In this paper, we present a simulation study on the performance of the optimal linear combination of longitudinal markers for binary disease classification under three missing data mechanisms, namely, missing completely at random, missing at random and non-ignorable missing.

APPLICATIONS IN BIOLOGY (Contributed, 4:15-6:00)

Multivariate Regression Tree: Classification of Bird Assemblages Based

on Their Habitat Characteristics

Marie-Héléne Ouellette, Université de Montréal, marie-helene.ouellette@umontreal.ca
Jean-Luc Desgranges, Service Canadien de la Faune, jean-luc.desgranges@ec.gc.ca
Pierre Legendre, Université de Montréal, pierre.legendre@umontreal.ca
Daniel Borcard, Université de Montréal, daniel.borcard@umontreal.ca

Abs: Ecological problems linked to water level fluctuations (for example due to dams) are numerous and often not very well known. The analyses presented in this paper come within a research program of the Commission mixte internationale de gestion des eaux des Grands Lacs et du Saint-Laurent (CMI). They revolve around bird assemblages along the Saint-Laurent river, one of the biggest water flow in Canada, in correlation with their habitat. The goal is to use these assemblages as bioindicators of the waterside surroundings ecological state. One of the techniques applied here consists in a multivariate regression tree, an ordination and clustering method developed by Death in 2002, focused on a selection of 128 sites. This tree is a model that permits us to predict either the birds' abundances in relation to the environmental characteristics, or the habitat characteristics in function of the birds abundances. A tree network allows us to find one of the best trees, explaining the most variance of the response table by exploring the effect of omitting certain variables from the modeling process. This one allows us to distinguish 6 groups of sites characterized by their specific bird assemblages and environmental properties.

Global Classification of (Plant) Proteins across Multiple Species

Naomi S. Altman, Pennsylvania State University, naomi@stat.psu.edu
Kerr Wall, Pennsylvania State University, pkw11@psu.edu
Jim Leebens-Mack, Pennsylvania State University, jhl10@psu.edu
Victor Albert, University of Oslo, victor.albert@nhm.uio.no

Abs: With rapidly growing numbers of whole genome and expressed sequence tag (EST) sequences in our public databases, sequence-based protein classification systems are providing foundations for gene annotation, functional genomics, and comparative investigations of gene and genome evolution We use the similarity-based clustering procedure TribeMCL (Enright et al 2002, 2003) to classify protein-coding genes into putative gene families for Arabidopsis, rice and poplar. The results of these analyses provide insights into the Arabidopsis, rice and poplar proteomes, protein family evolution, and the evolutionary dynamics of functional domains among gene families. Phylogenetic analyses of exemplar gene families shows a strong, but not perfect correspondence between tribe membership and cladistic relationships. One of the challenges of this type of classification is determining the stability of the proposed families under small changes in the data. Classifications have been constructed using three clustering stringencies and the strength of support for clusters as historical entities has been tested through jackknife analyses and bagging methodology.

Phyloinformatics of Genes in Complete Chloroplast Genomes

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

Abs: Molecular phylogeny is important in the study of related species based on structural similarity. Chloroplast genomes have been used to study important genes and to help in understanding important plant traits like disease resistance, herbicide resistance, and crop yield. Using model genes like Rubisco (ribulose-1,5-bisphosphate carboxylase) can help in the crucial analysis on phylogenetic reconstruction based on structural-activity across species. The conserved nature of this important gene has been studied using informatics tools that find motifs in orthologous sequences using rVISTA. In order to find weak motifs in the sequences used, PVLMM model is used. PVLMM (Permuted Variable Length Markov Model) is based on Variable Length Markov Model (VLMM) where we have a distribution for X_1, and, for $l = 2, \ldots, L$, we have a constrained conditional distribution for X_l given $X_{l-1}, \ldots, X_1$. In order to model PVLMM we need to introduce L-1 context functions. To interpret the selected PVLMM, sequence logos are used. Results indicate how statistical modeling can be used to visualize and identify dependency of motifs. This is especially important in conserved molecules for phylogenetic analysis in the evolutionary process of plant species.

Visualizing Primate Evolution: Reification of a Statistical Model

F. James Rohlf, Stony Brook University, rohlf@life.bio.sunysb.edu
Nina Amenta, University of California, Davis, amenta@cs.ucdavis.edu
Eric Delson, City University of New York, delson@amnh.org
David F. Wiley, University of California, Davis, wiley@cs.ucdavis.edu
Will Harcourt-Smith, City University of New York, willhs@amnh.org
Steve Frost, University of Oregon, fabiofrost@yahoo.com
Alfred L. Rosenberger, City University of New York, alfredr@brooklyn.cuny.edu
Dan A. Alcantara, University of California, Davis, dfalcantara@ucdavis.edu
Lissa Tallman, City University of New York, mtallman@nyc.rr.com

Abs: Evolutionary biologists are often concerned with identifying intermediate shapes in an evolutionary sequence or inferring, in a rigorous way, what a hypothetical ancestor might have looked like. Consider the past 150 years during which time scientists have pondered how the first members of the human lineage might have looked - after departing phylogenetically from an African ape ancestor. We are developing a set of tools to address this matter: how to mathematically constrain and visualize shape transformation based on knowledge from phylogenetics. As a case study we examine the evolution of Old World monkeys.

The locations of biological landmarks were identified on laser surface scans (LSS) of the skulls of five extant species of Old World monkey (Family Cercopithecidae). In addition, semi-landmark points were located to capture additional shape information along surfaces between the landmarks. The locations of the semilandmark points were slightly adjusted to minimize differences in shape resulting from arbitrary differences in the spacing of the semi-landmarks. Generalized Procrustes analysis was then used to align the configurations

of landmarks and to project each specimen into a multivariate space that is tangent to Kendall's shape space. This allows variation among shapes to be analyzed using conventional multivariate methods. A phylogeny was imbedded in this space so that the squared length of the implied trajectory was minimized. Points along this trajectory were then projected back into the physical space of the organism and visualized as 3-dimensional representations of the surfaces of hypothetical skulls which are estimates of ancestral forms along the evolutionary trajectories. The 3-dimensional representations are produced by warping all of the original surface scans so that their landmark points coincide with the landmark configuration corresponding to the point of the trajectory, and then merging the surface scans into a single surface.

These methods can also be extended to allow the visualization of shape variation implied by other types of multivariate analysis such as multivariate multiple regression and related methods.

ENVIRONMENTAL APPLICATIONS (Invited, 8:45-10:30)

Statistical Modeling and Evaluation of Microbial Source Tracking Data from rep-PCR DNA Fingerprints

Luis Tenorio, Colorado School of Mines, ltenorio@mines.edu
Junko Munakata-Marr, Colorado School of Mines, jmmarr@mines.edu
John Albert, Colorado School of Mines, jalbert@mines.edu

Abs: The goal of this talk is to outline statistical considerations for the development of source identification methods based on DNA fingerprinting. Microbial source tracking (MST) techniques are used to identify sources of fecal contamination. Here, we focus on a particular DNA fingerprinting technique based on the analysis of repetitive units in the genomic DNA via the repetitive element polymerase chain reaction (rep-PCR). This technique is widely used for MST. We report results of a reproducibility study of densitometric curves for source identification through library trained classification methods.

Simulation-Based Detection of Water-Borne Bacterial Contamination

Brian West, Applied Maths, brian_west@applied-maths.com

Abs: Tracking the source of water-borne bacterial contamination requires the demonstration of a non-random similarity between a source population and a contaminated population. The population similarity coefficient, Sp, is a statistical measure of the extent to which two populations have similar bacterial type frequencies. This allows non-random similarities to be detected even when there is significant overlap between uncontaminated populations. Assuming that Sp values vary continuously, a confidence value must also be calculated. One way to measure confidence is by iteratively re-sampling the populations with replacement, i.e. bootstrapping, and tabulating the likelihood that such a high Sp is obtained by chance. This has the advantages of deriving, rather than assuming, the distribution of Sp values, and of being conceptually straightforward. This method is tested on bacterial isolates from multiple source populations, using within-population comparisons and simulated contaminations as controls. The results are compared to those derived from more traditional measures of population similarity.

Scaling by Reference Conditions for Ecological Assessment

Samantha Bates Prins, Virginia Polytechnic Inst. and State University, sprins@vt.edu
Eric P. Smith, Virginia Polytechnic Institute and State University, epsmith@vt.edu

Abs: Reference sites provide important information about the range of biological, physical and chemical measurements. Using reference information to scale data from other sites is useful for evaluating the status of sites and establishing impairment. A common approach is to use all the data on the reference conditions to standardize measurements for a new site. Rather than using all available reference sites to scale the observed value of a particular

metric at a test site, we present an alternative approach that uses only the k closest (in terms of selected predictors) reference sites. These selected predictors may include only natural variables (for instance latitude and longitude); only potential stressors that naturally occur, or a combination of both, and these variables may be continuous or categorical. We believe this nearest neighbors based distribution of the metric at a test site is closer to the proper reference distribution for that site. Using a set of data from the Mid-Atlantic Highlands, we show that the nearest neighbor method improved on the ability of the regression approach to classify test sites correctly without affecting the ability to predict reference sites. The proposed approach was also more sensitive. We discuss issues related to the choice of k and present results suggesting an optimal k in this application.

MODEL BUILDING: MIXTURES & BIOINFORMATICS (Invited, 8:45-10:30)

Analysis of Semiparametric Mixture Models, with Application to QTL Analysis

Jason Fine, University of Wisconsin, Madison, `fine@biostat.wisc.edu`

Abs: In this talk, we propose a semiparametric alternative to traditional parametric QTL interval mapping. Our model assumes that the log ratio of the component densities satisfies a linear model, with the baseline density unspecified. We begin by considering the simple case of single gene model in backcross, with two component mixtures. We show that a constrained empirical likelihood has an irregularity when the two densities are equal. A partial empirical likelihood is proposed which permits unconstrained estimation of the parameters and gives consistent and asymptotically normal estimators. It turns out that the partial empirical likelihood is related to a conditional likelihood involving additional nuisance parameters. We establish that the partial likelihood estimator is more efficient than an estimator with the nuisance parameters known. The practical utility of the methods is illustrated on a rat study of breast cancer resistance genes. The approach is then extended to intercross designs and multiple gene models. unifying standard parametric multigene models, including epistasis.

A Mixture Model Approach to Multiple Hypothesis Testing

Geoff McLachlan, University of Queensland, `gjm@maths.uq.edu.au`
Liat Ben-Tovim Jones, University of Queensland, `liatj@maths.uq.edu.au`

Abs: With the statistical analysis of data sets today, there is increasing need to carry out tests on many hypotheses. For example, in bioinformatics, an important and common problem in microarray experiments is the detection of genes that are differentially expressed in a given number of classes. As this problem concerns the selection of significant genes from a large pool of candidate genes (possibly in the tens of thousands), it needs to be carried out within the framework of multiple hypothesis testing. In this talk, we focus on the use of mixture models to handle the multiplicity issue. With this approach, a measure of the local FDR (false discovery rate) is provided for each gene (feature variable), and it can be implemented so that the implied global FDR is bounded as with the Benjamini-Hochberg methodology based on tail areas. The latter procedure is too conservative, unless one modifies it according to the prior probability that a gene is not differentially expressed. An attractive feature of the mixture model approach is that it provides a framework for the estimation of this prior probability.

Searching High-Dimensional Spaces for Parsimonious Mixture Models

R. Pilla, Case Western Reserve University, candr@herine.net
Catherine Loader, Case Western Reserve University, candr@herine.net

Abs: We consider the problem of clustering data in high-dimensional spaces. Existing algorithms for maximum likelihood, such as the EM algorithm, can be particularly slow when there are a large number of poorly separated component clusters. In this work we combine ideas of estimation and model selection in building efficient clustering algorithms. For nearby clusters, or when mixture components are poorly separated, selection rules based on penalized information criteria are employed to determine the value of separate clusters, versus single large clusters. Constructing a traveling salesman path through candidate cluster centers or mixture components enables us to efficiently pair nearby clusters, thereby enabling us to rapidly prune the set of candidates. Rules for defining parameter distance, updating weights and locations, and decisions on merging and deleting nodes are discussed.

TREES AND NEURAL NETWORKS (Invited, 8:45-10:30)

Gaussian Process Trees

Robert B. Gramacy, University of California, Santa Cruz, rbgamacy@ams.ucsc.edu
Herbert K. H. Lee, University of California, Santa Cruz, herbie@ams.ucsc.edu

Abs: The Gaussian Process (GP) is a popular tool for non-parametric regression for many reasons. GPs are conceptually straightforward, easily accommodate prior knowledge in the form of covariance functions, and return a confidence around predictions. In spite of their simplicity, three important disadvantages to standard GPs greatly restrict the types of data to which they are applied. Firstly, inference on the GP scales poorly with the number of data points, typically requiring computing time that grows with the cube of the sample size. Secondly, GP models are usually stationary in that the same covariance structure is used throughout the entire input space. Finally, the error (standard deviation) associated with a predicted response under a GP model does not directly depend on any of the previously observed output responses. Even in geostatistical research and design of computer experiments— arguably the GP's two biggest customers— response surfaces can be highly non-stationary and subject to region-specific noise. Traditional stationary GPs are essentially useless in such applications. All of the above shortcomings may be addressed by partitioning the input space into regions, and fitting separate GPs within each region. Partitioning allows for the modeling of non-stationary behavior, and can ameliorate some of the computational demands by fitting models to less data. A fully Bayesian approach can yield a region-specific measure of uncertainty in predictive inference. This talk extends the work of Chipman et. al (2002) on Bayesian CART (Classification and Regression Trees). Bayesian treed GP modeling is outlined in detail and demonstrated on several synthetic and real-world non-stationary datasets.

Making Use of Small Samples for Classification

Russell Steele, McGill University, steele@math.mcgil.ca
Matthew Taddy, McGill University, taddy@math.mcgill.ca
Inti Zlobec, McGill University, inti.zlobec@mail.mcgill.ca
Nilima Nigam, McGill University, nigam@math.mcgill.ca

Abs: In this talk, we will discuss several classification techniques, ranging from classification trees to Bayesian neural networks, in the context of developing predictive biomarker models for the efficacy of colorectal cancer treatment. Although most classification approaches address problems with large numbers of features and observations, we will focus on what types of inference can be drawn from applying sophisticated models to small datasets. In general statistical practice, limited sample sizes indicate a possible gain through the use of Bayesian methods, but the choice of prior distributions and the estimation of posterior distributions can often be difficult to implement for relatively statistically sophisticated clinical researchers. The speaker will present admittedly ad hoc, but otherwise effective techniques that allow even relatively un-statistically sophisticated clinical researchers to make reasonable and useful inference.

Default Bayesian Neural Network Classification

Herbie Lee, University of California–Santa Cruz, herbie@ams.ucsc.edu

Abs: Neural networks can be highly effective for classification, but the parameters can be difficult or impossible to interpret. This presents problems in specifying a coherent prior under the Bayesian approach. This talk presents some methods for specifying non-informative priors, giving a way to formally quantify our ignorance about the parameters and allow the data to be maximally informative.

DATA MINING (Invited, 10:45-12:30)

Learning Classifiers Under a Limited Budget for Acquiring Training Data

Balaji Krishnapuram, Siemens Medical Solutions, Balaji.Krishnapuram@siemens.com

Abs: As compared to passively learning from the set of examples provided by a teacher, a student can learn concepts much faster if he actively asks questions to clarify the most confusing ideas. In an effort to minimize the data acquisition budget, this idea has been repeatedly exploited over the years (e.g., in the form of Bayesian design of experiments) to reduce the number of training samples required to learn accurate classifiers. During the last decade, another stream of exciting research has focused on addressing the data insufficiency problem using a complementary approach. In many situations, samples are first collected unlabeled, and due to budgetary constraints, class label information is subsequently collected only for a subset of the samples, possibly using active learning strategies. In semi-supervised learning, one uses both the labeled training set, as well as the unlabeled data in the process of estimating the classifier; if the cost of acquiring unlabeled data (i.e., features) is much less than the cost of acquiring class labels, this approach can significantly decrease the data acquisition cost. In this talk I will discuss another set of three closely related strategies that can also be used to address the problem of data insufficiency: (a) Co-training, where one enforces internal consistency on classifiers for different physical sensors; (b) Multi-task learning, where one exploits the statistical dependence between different classification problems and (c) Concept change where one uses historical data drawn from a distribution that is known to be slightly different from that of the test data that we want to classify. I will provide new algorithms that exploit these ideas; practical illustrations will be provided to study how these approaches work in real life classification problems.

Nonparametric Statistics and Data Mining

Tamraparni Dasu, AT&T Labs, `tamr@research.att.com`

Abs: A typical data mining task involves data sets that are massive, heterogeneous, with complex interdependencies between many variables. The true structure in the data is often muddied by hidden anomalies resulting from data quality issues. Parametric approaches tend to be too restrictive in their assumptions to generalize well to a wide variety of data mining applications. Instead, data driven approaches based on nonparametric statistical techniques are more effective and widely applicable. One such approach employs data partitions. Partitioning the data into meaningful classes and using class-based summaries for further analysis makes it possible to rapidly mine massive, multivariate data sets.

In this talk, we discuss four major types of partitioning schemes - 1) stratification, where a pre-defined set of criteria such as the levels of a categorical attribute, are used to define the partition, 2) induced partitions that are the result of an underlying model like clustering or classification, 3) database friendly schemes, such as data cubes that use the semantics of variables to create linear, often hierarchical classes and 4) general partition schemes based on depth equivalence classes. Each partitioning scheme has its advantages and disadvantages.

Once a partitioning scheme has been chosen, we need to compute a set of representative summaries for each class. These summaries collectively represent a profile of the data that can be used for further analysis. The summaries should have certain desirable characteristics: statistical properties such as unbiasedness, consistency, sufficiency and efficiency; computational properties such as computational speed, extensibility to higher dimensions and storage requirements; and finally ”aggregability” i.e. the ability to aggregate statistics from a finer partition to a coarser partition without accessing the original data. We conclude with important real-life applications of partition based data mining to telecommunication data.

Mining Temporal Patterns

Feng Liang, ISDS, Duke University, feng@stat.duke.edu

Abs: We consider the problem of discovering temporally dependent patterns with some time relationships. Previous work of temporal mining focuses on frequent itemsets with a predefined time window. The fixed time-window scheme will miss temporal relationships longer than the window size and significant, but infrequent patterns. Aiming to tackle the two problems, we propose an algorithm for discovering temporal patterns without predefined time windows. The problem of discovering temporal patterns is divided into two sub-tasks: (1) using “cheap statistics” for dependence testing and candidates removal (2) identifying the temporal relationships between dependent event types. The dependence problem is formulated as the problem of comparing two probability distributions and is solved using a technique reminiscent of the distance methods used in the spatial point process, while the latter problem is solved using an approach based on Chi-Squared tests. The statistical properties provide meaningful characterizations for the patterns and are usually robust against noise. The algorithms are applied to the problem of mining patterns in real data collected from production networks, producing some interesting insights.

DEVELOPMENTS IN BIOINFORMATICS (Invited, 10:45-12:30)

Selecting an Appropriate Clustering Algorithm for Analyzing Microarray Data

Susmita Datta, Georgia State University, sdatta@mathstat.gsu.edu

Abs: Cluster analysis is most commonly used to group thousands of genes on the basis of their similarity in expression profiles in a microarray experiment. However, it has been noted that the genes are grouped differently for different clustering algorithms. Hence, selecting a clustering algorithm that is optimal in some way from a rather impressive list of clustering algorithms that currently exist is a challenging problem. To address this problem most of the approaches rely on the quality of the clusters produced in terms of their physical characteristics and statistical properties such as stability or consistency. While this may be reasonable for certain applications such as machine learning, the underlying biology must be taken into consideration for applications to microarray experiments. There have been attempts in recent years in interpreting the resulting clusters obtained using a statistical clustering method in terms of their biological functions. However, the problem of validation and selection of a clustering algorithm on the basis of the functional information has received relatively little attention. Here in this research we incorporate both statistical stability and functional validity in selecting an optimal clustering algorithm for a microarray data.

Estimating Network Topology and Latent Factors in A Gene Network Based on Independent Component Analysis

Wei-Fuh Wang, National Taiwan University, gshieh@stat.sinica.edu.tw
Chung-Ming Chen, National Taiwan University

Abs: Latent factors, though not observable from the experiments, are generally of existence and potentially play influential roles on the gene expressions of interest. While various approaches have been proposed to estimate the genetic network from a set of microarray data, the influence of the latent factors has not been taken into account by most of previous reconstruction algorithms. As a consequence, the gene-gene interactions may be over- or under-estimated, even with a correct network topology. To account for the effects of latent factors, a new gene network reconstruction algorithm based on independent component analysis is proposed in this paper. The expression level of each gene is assumed to be a linear function of latent factors and observable gene expressions. The latent factors are extracted from the observed gene expressions by using independent component analysis, the cost function of which is the negentropy of the latent signals regularized by the mean squared errors between the predicted and observed gene expressions. Bayesian information criterion (BIC) and Akaike information criterion (AIC) are used as the cost functions to determine the number of latent factors and the network topology. The optimal solutions are sought by simulated annealing incorporating Newton-Raphson method. To evaluate the performance, the proposed algorithm has been validated by using simulated time-course data and a signal network controlling the switch from the mitotic cell cycle to meiosis in yeast.

Statistical Methods to Construct Transcriptional Regulatory Networks Using Gene Expression and DNA Sequence Data

Biao Xing, Genentech, Inc, xing.biao@gene.com
Mark van der Laan, University of California–Berkeley

Abs: Transcriptional regulatory networks specify regulatory interactions among regulatory genes and between regulatory genes and their target genes. Uncovering transcriptional regulatory networks helps us to better understand complex cellular processes and responses. We present two statistical methods for constructing transcriptional regulatory networks using gene expression data, promoter sequences, and transcription factor binding sites. Both methods start from identifying active transcription factors under each individual experiment, using a feature selection approach. The first method employs a naive normal mixture model to classify the transformed gene expression data for each transcription factor and uses the posterior probability of being in the 'induced' or 'repressed' classes to measure the strength of regulatory interactions. Evidence is averaged across different experiments to infer the overall regulatory network structures. The second method employs a causal inference model to model the causal effect of a transcription factor on its potential target genes. A nonparametric marginal structural model is built for every transcription factor and gene pair, which also allows controlling for potential confounding effects of other transcription factors on the expression level of the gene. The p-value associated with the causal parameter in each of these models is used to measure the regulatory interaction strength. These results are used to infer the overall regulatory interaction matrix and network structures. Simulation studies and analysis of yeast data have shown that both methods are capable of identifying significant transcriptional regulatory interactions and uncovering underlying regulatory network structures and both can be complementary to each other to maximize significant finding.

AUTHOR IDENTIFICATION (Invited, 10:45-12:30)

Lexical Predictors of Personality Type

Shlomo Argamon, Illinois Institute of Technology, argamon@iit.edu

Abs: We are currently pursuing methods for ”author profiling” in which various aspects of the author’s identity might be identified from a text, without necessarily having a corpus of documents from the same individual. A key component of such an identity profile is personality; this talk will address distinguishing high from low neuroticism and extraversion in authors of informal text. We consider four different sets of lexical features for this task: a standard function word list, conjunctive phrases, modality indicators, and appraisal adjectives and modifiers. SMO, a support vector machine learner, was used to learn linear separators for the high and low classes in each of the two tasks. We find that appraisal use is the best predictor for neuroticism, and that function words work best for extraversion. Further, examination of the specifically most important features yields insight into how neuroticism and extraversion differentially affect language use. [Joint with M. Koppel and J. Pennebaker]

Multiple Methods and the Federalist Papers

Ross Sowell, University of the South, rsowell@gmail.com
Diana Michalek, University of California–Berkeley, dianam@gmail.com

Abs: Problems of authorship identification have caused controversy in areas such as history, literature, and forensics for many years. One well-known example is that of the disputed Federalist Papers. In 1964, Frederick Mosteller and David Wallace used thirty function words and Bayesian inference to attribute all the disputed papers to Madison. In this paper, we use the frequency of Mosteller and Wallace function words as well as word length frequencies to classify the papers. We present a simple method for ranking these features

according to their discriminating ability, and we show that by using the top ranked feature from each list, we are able to use linear discriminant analysis to classify all the disputed papers to be of Madisonian authorship.

Author Identification on the Large Scale

Alex Genkin, DIMACS, Rutgers, kantorp@research.rutgers.edu
D. Lewis, Consultant, ddlewis2@worldnet.att.net
D. Madigan, Rutgers, madigan@stat.rutgers.edu
D. Fradkin, Rutgers, dfradkin@dimacs.rutgers.edu

Abs: Author attribution has a long history that includes some famous disputed authorship cases and also has forensic applications. The focus of this research is on large collections of documents, large groups of potential authors (hundreds and thousands), and thousands of stylometric features. We chose Bayesian polytomous regression as our statistical tool and developed software that handles problems of that size, now publicly available. We experimentally studied different sets of stylometric features and their combinations and found their relative value for classification to be pretty stable on different data collections. We will also talk about how to separate author identification from topic categorization.