CSNA-2007 Abstracts (Ordered alphabetically by first author's name)

The 2007 Meeting of the Classification Society of North America is Sponsored by National Center for Supercomputing Applications, the UIUC College of Education, Department of Computer Science, and the Graduate School of Library and Information Science. Special thanks also to CRC Press and to Wolfram Research.

How are bullies and victims embedded in the classroom peer network? : Factors linked to the heterogeneity of their structural embeddedness

Hai-Jeong Ahn, Claire Garandeau

University of Illinois at Urbana Champaign

Abs: We examined the structural embeddedness of bullies and victims in the classroom peer network. 3rd and 4th graders (N = 680) completed within-class peer nominations for bullies, victims, and behavioral profiles (i.e. aggressive and prosocial behavior). 147 bullies (78 boys) and 150 victims (77 boys) were identified after controlling for gender and class. Classroom social network structure was determined by cohesive blocking procedure based on affiliation relationships (i.e. hang around together). Classroom social network embeddedness varied across classes. The structural embeddedness of bullies and victims was heterogeneous. 35% of bullies and 32% of victims are deeply embedded in the classroom social network, but 22% of bullies and 30% of victims are placed outside of the social network. Boy bullies are more deeply embedded in the peer network than girl bullies, but no gender differences are found for victims. Black bullies and victims are more deeply embedded in the classro! om! social network structure than non-black bullies and victims. We analyze classroom racial composition and behavioral profiles of bullies and victims in order to clarify the heterogeneity of their structural embeddedness.

Dimension Reduction Using a Hybrid of PARAMAP and Isomap Procedures

Ulas Akkucuk, Bogazici University, Istanbul, Turkey

J. Douglas Carroll, Rutgers Business School

Abs: Dimensionality reduction aims to represent higher dimensional data by a lower-dimensional structure. A well-known approach by Carroll, Parametric Mapping (abbreviated PARAMAP) relies on iterative minimization of a loss function measuring the smoothness or "continuity" of the mapping from the lower dimensional representation to the original data. This approach was primarily only of theoretical interest at the time it was developed, since it was computationally expensive and prone to local optima. The approach was resuscitated recently with important algorithmic modifications and we will call this new approach "PARAMAP-2". In the earlier PARAMAP approach, when given data sets that either were "large", had added error, irregular spacing, or all of these features combined, the method tended essentially always to converge to a clear local optimum, and was incapable of obtaining the globally optimal solution known to exist. Using the PARAMAP-2 approach we were able to obtain solutions we were certain were globally optimal in reasonable computational time. In this paper we discuss use of a variant of a method called Isomap to obtain a starting framework, and then adding new points in batches based on their proximity to landmark points in this initial framework using the PARAMAP-2 algorithm. Since Isomap is faster and less prone to local optimum problems than PARAMAP, and the iterative process involved in adding new points to the configuration should be much less time consuming we believe the resulting method should be much better suited to dealing with large sets of realistically based data, and more inclined to obtain a satisfactory solution in roughly order n^2 time. One initial test of a version of this new "PARAMAP-3" algorithm has been implemented, with very promising results. In this paper we will explain the methods of obtaining the framework and the iterative procedure of mapping in the rest of the points, demonstrating this methodology on the well-known nonlinear manifolds such as the sphere and four dimensional torus.

On Similarity Indices and Correction for Chance Agreement

Ahmed N. Albatineh, Nova Southeastern University
Magdalena Niewiadomska-Bugaj, Western Michigan University
Daniel P. Mihalko, Western Michigan University

Abs: Similarity indices can be used to compare partitions (clusterings) of a data set. Many such indices were introduced in the literature over the years. We are showing that out of 28 indices we were able to track, there are 22 different ones. Even though their values differ for the same clusterings compared, after correcting for agreement attributed to chance only, their values become similar and some of them even become equivalent. Consequently, the problem of choice of the index to be used for comparing different clusterings become less important.

An ordinal impurity function for classification trees when predicting an ordinal response

Kellie J. Archer, Department of Biostatistics, Virginia Commonwealth University

Abs: Ensemble methods have been demonstrated to be competitive with other machine learning approaches for classification and have been described for nominal, continuous, and survival responses. However, in a large number of biomedical applications, the class to be predicted may be inherently ordinal. Examples of ordinal responses include TNM stage (I, II, III, IV) and drug toxicity (none, mild, moderate, severe). While nominal response methods may be applied to ordinal response data, in so doing some information is lost that may improve the predictive performance of the classifier. This study examined the effectiveness of using both an ordinal impurity function for classification tree growing and bootstrap aggregation. Results using the ordinal impurity function are compared to those obtained using the Gini impurity function, Gini impurity with a linear loss, and Gini impurity with quadratic loss on both simulated and benchmark datasets.

Influence analysis in Depth Transvariation based Classification

Nedret Billor, Asheber Abebe, Asuman Turkmen, and Sai Nudurapati,
Department of Mathematics and Statistics, Auburn University, AL

Abs: Depth transvariation based classification is a new nonparametric classification technique based on classifying a multivariate data point through maximizing the estimated transvariation probability of statistical depths.

In this paper we will study the influence of observations on the misclassification error rate in depth transvariation based discriminant analysis. We assess the partial influence of the error rate , which allows us to quantify the effect of observations in the training sample, based on the performance of the depth transvariation based classification rule.

We use some simulated data sets as well as some real examples to evaluate the robustness of error rate based on the depth transvariation classification rule.

Challenges in the classification of spatial data

Alexander Brenning, Department of Geography, University of Waterloo, Ontario, Canada

Abs: While the amount of spatial data available within Geographical Information Systems is exploding, research on the construction and assessment of classification techniques for spatial data is still very limited. In the case of gridded spatial data such as remote-sensing data or landslide inventories, one common characteristic that neighboring raster cells have to be considered pseudo-replications representing virtually identical observations. This phenomenon, and more generally the presence of spatial autocorrelation, has important implications: (1) Successful classifiers must avoid over-fitting in the presence of spatial autocorrelation, and (2) appropriate techniques for the estimation of misclassification error rates, such as spatial cross-validation or a spatial bootstrap, are needed for error assessment. Results of classification and estimation techniques in two different applications are presented: The two-class problem of landslide susceptibility mapping, and the multi-class problem of crop identification from multi-temporal remote-sensing data.

Title: Classification of Massive, Structured Data: Research Progress @ Data Mining Group.CS.UIUC

Deng Cai, University of Illinois

Abs: During the past decades, numerous classification algorithms have been proposed. Many of them take the numerical feature vectors as input and predict the label of a sample. However, much real-world data is not represented as numerical vectors, but as more complicated structures, such as sequences, trees, or graphs. To solve these real-world problems, we need to develop more flexible classification based approaches. Our group aims at developing algorithms that can well handle the real-world problems. Specifically, we are interested in and developing algorithms for the following problems:

Frequent pattern based classification The frequent pattern (sub-structure) in complicated structured data can naturally be used as features. How to efficiently mining these sub-structures? How to select the most powerful substructures with respective classification performance? We conducted a systematic exploration of frequent pattern-based classification, and provide solid reasons supporting this methodology.

Name entity distinction Different people or objects may share identical names in the real world, which causes confusion in many applications. It is a nontrivial task to distinguish those objects, especially when there is only very limited information associated with each of them. Although linkages among objects provide useful information, such information is often intertwined and inconsistent with the identities of objects. Moreover, different types of linkages carry different semantic meanings and have different levels of pertinence. We developed a general object distinction methodology called DISTINCT for analysis based on supervised composition of heterogeneous links.

Classification on Stream In recent years, there have been some interesting studies on predictive modeling in data streams. However, most such studies assume relatively balanced and stable data streams but cannot handle well rather skewed (e.g., few positives but lots of negatives) and stochastic distributions, which are typical in many data stream applications. We proposed a new approach to mine data streams by estimating reliable posterior probabilities using an ensemble of models to match the distribution over under-samples of negatives and repeated samples of positives.

Trajectory classification With the emerging GPS and RFID technologies, modeling trajectories on road networks becomes more and more important for transportation and traffic planning. We studied methods for classifying trajectories on road networks. By analyzing the behavior of trajectories on road networks, we observe that, in addition to the locations where vehicles have visited, the order of these visited locations is crucial for improving classification accuracy. Based on our analysis, we contend that frequent sequential patterns are the best feature candidates since they preserve this order information. Furthermore, when mining sequential patterns, we propose to confine the length of sequential patterns to ensure high efficiency. Our comparative study over a broad range of classification approaches demonstrates that our frequent pattern-based classification method significantly improves accuracy over a naïve method in some synthetic trajectory data.

Automatic Software Debug Automated localization of software bugs is one of the essential issues in debugging aids. What can be the effective features for automatic software debugging? What statistical model is suitable for this problem? We have proposed a new statistical model-based approach, called SOBER, which localizes software bugs without any prior knowledge of program semantics.

Discriminant Analysis for Large Scale High Dimensional Data Many real-world data are with high dimensionality, so as the generated features (frequent patterns) of structured data. When one performs discriminant analysis on such data set, the computational cost is always a bottleneck. Specifically, the most well applied algorithm, Linear Discriminant Analysis (LDA), has cubic-time complexity with respect to min(m,n), where m is the number of samples and n is the number of features. When both m and n are large, it is infeasible to apply LDA. We developed a novel algorithm for discriminant analysis, called Spectral Regression Discriminant Analysis (SRDA). SRDA has linear-time complexity with respect to both m and n.

Modularity Clustering for Thematic Document Clustering

Brant Chee and Bruce Schatz

University Of Illinois, Institute for Genomic Biology

Abs: We present a modified physics algorithm that is single link in nature that takes advantage of the inherent scale free and small world characteristics of word graphs in order to create semantic concept clusters of words. The word clusters are then used to segment large number of documents into overlapping thematic document clusters. Our method differs from existing methods such as Carrot 2 or latent semantic indexing in that the running time is roughly linear (O(n log2 n) in the number of terms). We have demonstrated results on large numbers of documents (¿100K) and faster running time than complete link algorithms while the clusters are more coherent and have higher utility than faster methods such as K-means.

Our algorithm takes advantage of the inherent clustering that exists in a small world graph whereas, the original physics algorithm makes assumptions about a log distribution in the number of elements in the in a cluster. We make no such assumptions and explicitly remove restrictions so that the number of elements in a cluster follows a more normal distribution providing better thematic document mappings.

Connections between K-Means Cluster Analysis and Restricted Latent Class Models

Chia-Yi Chiu, Department of Educational Psychology, University of Illinois

Jeff Douglas, Department of Statistics, University of Illinois

Abs: Restricted latent class models have seen many applications in psychometrics in recent years. One application has been in multiple classification latent class models where presence or absence of each of several latent attributes is under diagnosis. Maximum likelihood estimation or Bayesian computation, such as Markov chain Monte Carlo, are usually used to estimate parameters of these latent class models. K-means cluster analysis, when informed by the assumed structure of the latent class model, can be used as an alternative. By selecting an appropriate multidimensional statistic as an input of K-means, K-means results in almost identical classification as the conventional estimation methods can do, but does not need to access the complex models. This study will demonstrate how to select a statistic for K-means with latent class models and compares the results of the K-means clustering method with the conventional estimation method. An application to skills diagnosis with education testing data is presented.

Evaluation of Hierarchies based on the Longest Common Prefix, or Baire, Metric

Pedro Contreras and Fionn Murtagh Department of Computer Science

Royal Hollow, University of London

Abs: The Baire space has a metric that can be defined from the longest common prefix of two strings. Consider two floating point numbers with the first p digits identical. Then what we call their Baire distance is 2^{-p}. This distance is an ultrametric. It follows that a hierarchy can be used to represent the relationships associated with this distance. We address the issue of whether such a hierarchy (let us call it a Baire hierarchy, because it is clearly a different hierarchy compared to one resulting from any of the commonly used agglomerative hierarchical clustering algorithms) is advantageous, computationally, for clustering large, high dimensional data sets; and also how useful it is, in particular compared to k-means.

Putting Some 'WoW' into Modeling Longitudinal Networks

Bethany Dohleman, Department of Psychology, University of Illinois at Urbana-Champaign

Harold D. Green Jr., Science of Networks in Communities (SONIC), National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

Dmitri Williams, Department of Speech Communication, University of Illinois at Urbana-Champaign

Noshir Contractor, Department of Speech Communication & Science of Networks in Communities (SONIC), National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

Abs: This study concerns co-evolution of communication networks and perceived expertise among members of a Massively Multiplayer Online Role-Playing Game called World of Warcraft (WoW). We explored multi-theoretical multilevel hypotheses about what motivates an individual to create expertise-seeking advice ties from other players within their virtual communities, called guilds.

We used SIENA to test how co-evolutionary dynamics vary for guilds that have audio and text communication versus only text communication. Nine guilds received a Voice over Internet Protocol (VoIP) headset, allowing them to communicate with each other using audio and text means. Changes in network structure and member attributes in these treatment guilds are compared with changes in seven guilds that relied only on text means.

Results reveal that mechanisms associated with the co-evolution of net-

work structure and perceived expertise levels varied between treatment and control guilds. For example, experts in VoIP guilds are more likely to identify and approach other experts for advice. Other results reveal that for VoIP guilds, rates of communication are larger in the initial time interval than in other time intervals, a pattern not found in control guilds. Comparing patterns of co-evolution for control and treatment guilds demonstrates the application of new statistical tools for comparing parameters.

One-sided Elasticities and Technical Efficiency in Multi-output Production: A Theoretical Framework

Petros Hadjicostas, Department of Mathematics and Statistics, Texas Tech University (Joint work with Andreas C. Soteriou)

Abs: One of the concepts that have sparked considerable interest in the theory of production and efficiency is that of returns to scale (RTS). Economics researchers typically define RTS using the notion of elasticity. Considerable research activity on RTS has also been observed by management science researchers, who utilize the methodology of Data Envelopment Analysis (DEA) to gain insights on RTS. In this talk, we present a theoretical framework (developed jointly with Andreas Soteriou) that integrates existing economics and management science literature on RTS, and provides a foundation for research work in this area. Our framework defines, discusses, and proposes an approach to measure input- and output-oriented elasticities, and one-sided RTS. We demonstrate how the work done in DEA is a special case of our framework, and discuss the conditions under which the resulting two left-hand, and the two right-hand elasticities can be equal. The results of this work have been published in 2006 in the European Journal of Operational Society.

Current and future research directions are also discussed. For example, in a recent working paper, Hadjicostas and Soteriou explore properties of different orders of one-sided elasticities in multi-input multi-output production using the aforementioned theoretical framework. A special case of the theory in the paper is the Banker-Morey (1986) DEA model for data that include both discretionary and non-discretionary inputs and outputs.

A Useful Combinatorial Lemma and Some Applications

Bernard Harris

Abs: About 50 years ago, Leo Katz introduced a formula that he used to determine the probability that a random mapping is connected. This formula is revisited and generalized. Several applications of the original result and the generalization are discussed.

Identification of Multiple Functional Peaks resulting from a Common Peak Shape Function

Matt Hersh, Kert Viele, and Robin Cooper

Departments of Statistics and Biology, University of Kentucky

Abs: In synaptic transmission data, we observe electrical currents across time. These electrical current traces have peaks resulting from the release of transmitter from a vesicle across the synaptic cleft. If multiple vesicles release transmitter, one may observe multiple peaks within the electrical current trace. Previous work indicates that the electrical current from a single vesicle release, while varying in peak intensity and width of the synaptic potential, follow a similar shape. Thus, we analyse our electrical currents under the assumption that each function is the sum of equally shaped (though differingly scaled) peak functions. This has similarities to a mixture model, although instead of classifying individual observations into groups we only have the functional data. Our goals are to simultaneously estimate the underlying shape function and to classify each trace as to the number of peaks it contains. We use a mixture of functions to accomplish this task.

Intrusion Detection and Response using Effective Data Mining Techniques: Classification, Clustering and Data Analysis in Intrusion Information Retrieval

Dr. Emmanuel Hooper, Information Security Group, University of London

Abs: There are major challenges in intrusion detection and data mining, specifically, classification, clustering and data analysis of intrusion and alert datasets. Information retrieval in Intrusion datasets presents the problems of relevant feature attributes for effective detection and response to intrusions/attacks and alerts. Furthermore, the problem of false positives and detection capability of intrusion detection systems (IDSs) are major issues. This paper examines classification and clustering techniques for effective detection and response strategies to network infrastructure attacks. The approaches are two-fold: first, Classification of known attacks and alerts, and secondly, clustering of unknown attacks and alerts. The classification approach involves a hybrid of Bayesian Classification and Discriminant Analysis for known attacks and alerts. The clustering approach involves a hybrid of Ward's agglomerative algorithm and Pearson's Association Correlation along with Chi-square Analysis for unknown alerts and attacks. Abstract subcategories of the feature attributes are used to identify unknown attacks or benign alerts. Then appropriate responses are sent to mitigate the impact of the attacks and filter benign alerts from the IDS monitor. These strategies improve the performance of the IDS and enhance responses to various subcategories of false positives and complex attacks.

Text Classification with Customized Word Lists: Delta-Lz and Delta-Oz

David L. Hoover, New York University

Abs: J. F. Burrows recently introduced Delta, a new measure of textual difference for authorship attribution, and I have introduced five variants that improve upon Delta's already impressive results. Further testing confirms that two of the variants, Delta-Lz and Delta-Oz, are especially effective. My presentation reports the results of further inter-authorial and intra-authorial tests and investigates their implications. Both measures compare test and authorial text samples using individually determined and unique subsets of the word-frequency list for all test and authorial samples combined. Delta-Lz focuses only on words with test-sample frequencies quite different from the mean for the authorial samples: each test sample determines its own word-list. Delta-Oz compares each word's frequency in each test and authorial sample to its mean frequency in all the authorial samples. It compares only the frequencies of words with test-sample and authorial-sample frequencies that differ in opposite directions from the mean: each comparison between a test sample and an authorial sample determines its own word-list. Closely examining the words selected by Delta-Lz and Delta-Oz allows for a more accurate and informative characterization of the texts, and comparing the various lists should reveal why and how these methods are so effective.

Constructing A Music Mood Taxonomy By Clustering

Xiao Hu and J. Stephen Downie

GSLIS, University of Illinois at Urbana-Champaign

Abs: Classifying music pieces by the moods they express has attracted researchers' interests and efforts, but a standardized mood taxonomy is still in need for cross algorithm comparison and evaluation. This research strives to construct a reasonable music mood taxonomy for a common testbed of music mood classification in this year's Music Information Retrieval Evaluation eXchange (MIREX). We clustered the 40 most popular mood labels on a leading music website, AllMusicGuild (AMG) where each mood label is associated with a list of representative "Top Albums" and a list of "Top Songs". As these top albums and songs are often shared by different mood labels, they can be exploited to group the mood labels into several superclasses, which then would be a good candidate for a standardized mood taxonomy.

A co-occurrence matrix was formed for shared albums and shared songs respectively. Pearson's correlation was calculated for each pair of rows (or columns) as similarity measures between any two mood labels. Then a hierarchical clustering procedure using Ward's link was applied to such similarity measures. Comparing the clusters resulted from mood-album co-occurrences and mood-song co-occurrences, we found 29 mood labels were consistently grouped into 5 different clusters at about 1.5 distance level.

Avoiding Degeneracy in Multidimensional Unfolding by Combinatorial Optimization

Hans-Friedrich Köhn, Department of Psychology, University of Illinois, Urbana-Champaign

Abs: The recently introduced PREFSCAL algorithm (implemented in SPSS Categories) for fitting distance-based nonmetric unfolding models avoids degenerate solutions through augmenting the loss function by a penalty term to maintain a sufficient level of variability among the transformed proximities or pseudo-distances that prevents the distance estimates and associated object coordinates from collapsing into a single or a small number of locations.

We demonstrate that the same result can be obtained through a discrete (combinatorial) optimization strategy that does not involve any penalty functions for minimizing the loss function. In fact, PREFSCAL and combinatorial unfolding yield almost indistinguishable solutions for the test data sets employed.

The Astronomical Information Network

Michael J. Kurtz

Harvard-Smithsonian Center for Astrophysics

Abs: Astronomical Objects (stars, galaxies, ...) are bound together by a complex and often surprising assortment of shared ond/or similar interactions and histories. As a purely observational science it is the task of astronomy to disentangle the vast network of objects with shared or similar properties and discover the underlying causal relationships which govern our universe.

As an example some galaxies are blue, and exibit spectral features typical of a hot, ionized gas; other galaxies are red, and show none of these gaseous features. These red galaxies tend, very strongly, to cluster together in space, while the blue galaxies do not. Edwin Hubble first noticed this 75 years ago. What causes this effect, were the galaxies formed this way, or did they become this way over the history of the universe?

Astronomy research exists now within a large and growing man-made network of tightly interconnected data sources and services. Based on a culture of freely shared information and shared goals astronomers are building a Virtual Observatory, using the internet to bring the totality of astronomical information to anyone, anywhere in the world.

Astronomy research also exists within more abstract networks of thought and behavior. The structure of astronomy research can be seen in the co-citation network of astronomy research articles, but it can also be seen in the co-reader network of those articles, and the co-keyword network of those same articles again. Are these structures the same?

Logistic Regression using Fractional Imputation for Missing Data

Michael D. Larsen, Department of Statistics and Center for Survey Statistics and Methodology, Iowa State University

Abs: Imputation is used to fill in missing values so that analyses based on complete data methods can be completed. Random imputation methods can add imputation variance to the results. Not accounting for the fact that some records are imputed can lead to understatement of uncertainty in conclusions. A fractional imputation method for a missing outcome variable in logistic regression are proposed and implemented. The purpose of the methods is to reduce imputation variance while allowing accurate estimation of uncertainty. The method is applied to data from a longitudinal study of families in Iowa.

Non-Parametric Modeling of Partially Ranked Data

Guy Lebanon, Purdue University

Abs: Statistical models on full and partial ranking of n items are often of limited practical use for large n due to computational consideration. We explore the use of non-parametric models for partially ranked data and derive efficient procedures for their use for large n. The derivations are largely possible through combinatorial and algebraic manipulations based on the lattice of partial rankings. In particular, we demonstrate for the first time a non-parametric coherent and consistent model capable of efficiently aggregating partially ranked data of different types.

Analysis of Information Features in Natural Language Queries Seeking Music

Jin Ha Lee, University of Illinois at Urbana-Champaign

Abs: The deficiency of a formal taxonomy for representing real-life queries is a major barrier in appropriate design and evaluation of Music Information Retrieval (MIR)/Music Digital Libraries (MDL) systems. To address this issue, prior studies have analyzed real-life examples of natural language queries describing users' sought music and identified the general types of information features provided by users in their queries. However, the information regarding the empirical association among query features that would allow us to establish a meaningful classification scheme for these features is still lacking. This study aims to discover associative patterns in the kinds of information features provided in music queries by an empirical investigation. Real-life queries were collected from an online reference website and the categories of different information features were established by an iterative coding process. Different clustering approaches will be explored to infer associations from patterns of co-occurrences among these features and establish a classification scheme.

Large Scale Functional Data Clustering

Ping Ma, Department of Statistics, University of Illinois at Urbana-Champaign

Abs: Large scale functional data rise from many scientific investigations. Identifying cluster information is a crucial first step to navigate further scientific investigation.

Motivated by analyzing temporal gene expression data, we propose a novel functional data clustering method based on a mixture smoothing-spline model. For each cluster, we model its mean profile using a smoothing spline and describe its individual gene's variation by a parametric random effect. We present an EM algorithm to find the maximum a posteriori. Our method automatically takes care of the missing data and infers the number of clusters in the data. Emprical studies suggest that the proposed method outperforms the existing methods.

How Many Clusters?

Peter McCullagh, University of Chicago
Jie Yang, University of Illinois at Chicago

Abs: The title poses a deceptively simple question that must be addressed by any statistical model or computational algorithm for the clustering of points. Two distinct interpretations are possible, one connected with the number of clusters in the sample and one with the number in the population. Under suitable conditions, these questions may have essentially the same answer, but it is logically possible for one answer to be finite and the other infinite. This paper reformulates the standard Dirichlet allocation model as a cluster process in such a way that these and related questions can be addressed directly. Our conclusion is that the data are sometimes informative for clustering points in the sample, but they seldom contain much information about parameters such as the number of clusters in the population.

Visualizing Clusters With a Density-Based Similarity Measure

Rebecca Nugent, Department of Statistics, Carnegie Mellon University
Werner Stuetzle, Department of Statistics, University of Washington
Xiaoyi Fei, Department of Statistics, Carnegie Mellon University

Abs: The goal of clustering is to identify distinct groups in a dataset and assign a group label to each observation. To cast clustering as a statistical problem, we regard the data as a sample from an unknown density p(x). To generate clusters, we estimate the properties of p(x) either with parametric (model-based) or nonparametric methods. In contrast, the algorithmic approach to clustering (linkage methods, spectral clustering) applies an algorithm, often based on a distance measure, to data in m- dimensional space.

Many commonly used clustering methods employ functions of Euclidean distance between observations to determine groupings. Spherical groups are easily identified, curvilinear groups less so. We first motivate the use of a density-based similarity measure and briefly introduce generalized single linkage, a graph-based clustering approach. We describe a refinement algorithm used to bound the measure and then explore the performance of this measure in clustering and visualization methods.

Discrimination and Classification of Non-Stationary Brain Signals Using Higher Order Spectral Analysis

Hernando Ombao, University of Illinois at Urbana-Champaign
Ringo Ho, Nanyang Technical University

Abs: We consider a data set that consists of MEG (magnetoencephalogram) signals recorded from healthy controls and schizophrenic patients. Our neuroscience collaborators are interested in identifying features that can separate the two groups with the hope that these physiological measures may be used in conjunction with behavioral measures for patient diagnosis. In this talk, we will develop an automatic procedure for time-frequency spectral feature selection via localized transforms. Moreover, given the high degree of complexity of brain signals, we will consider the time- evolutionary spectrum of non-linear transforms as potential features for classification and discrimination.

An evaluation of social cognitive mapping procedures for identifying middle childhood social networks

Philip C. Rodkin and Hai-Jeong Ahn
University of Illinois at Urbana-Champaign

Abs: This study compares middle childhood social networks constructed from children's reports of: (a) their own affiliations and the affiliations of their peers (multi-informant affiliations); (b) their own affiliations (self-reported affiliations), and; (c) their own friendships. The sample consisted of 390 fourth and fifth graders surveyed in fall and spring. Multi-informant affiliation networks yielded larger peer groups and fewer isolates than networks constructed from friendship self-reports. Multi-informant affiliation networks were more stable from fall to spring and more robust to variations in statistical algorithms than friendship networks. Multi-informant affiliation and friendship networks had poor agreement with one another, particularly in their placement of unpopular children. Multi-informant affiliation groups had greater homophily on aggression than friendship groups and accorded higher social centrality to aggressive behavior. Self-reported affiliation networks were intermediate between multi-informant affiliations and self-reported friendships. Discussion focuses on incorporating multi-informancy into emerging sociometric technologies.

Racially integrated social networks among African- and European-American elementary children across differing classroom contexts

Philip C. Rodkin and Travis Wilson

University of Illinois at Urbana-Champaign

Abs: The racial composition of school classrooms, and the possible benefits of diverse classrooms, receives much attention in debates about school choice. We examine social integration between African- and European-American children across three elementary classroom contexts: 11 classrooms that are majority White (65% W, 35% B), 11 that are majority Black (65% B, 35% W), and 11 that are multicultural, with equal proportions of Whites and Blacks and 30% Hispanics and Asians. Participants were 680 3rd and 4th graders. Social integration was assessed by using a compositionally invariant ratio of same-race to cross-race nominations of who children: (a) affiliated with, (b) were friends with, (c) liked most, and (d) liked least. We expected interactions where the salience of classroom context to social integration would vary for European- and African-American children. Using multilevel modeling, we consistently found interactions between individual race and classroom racial majority. In majority white classrooms, African-American children had more segregated social relations and European- and African-American children mutually nominated one another as disliked. Along with substantive implications, discussion will focus on analytic concerns of how to classify children into groups, and how integrate metrics from research in child psychology and network science.

Objective Measurement of Fatigue in HIV/AIDS Using Actigraphy and Functional Data Analysis

William Shannon

Washington University School of Medicine

Abs: Behavioral changes associated with chronic HIV infection include lethargy or fatigue defined as the inability to continue functioning at a prescribed work rate in the presence of an increased perception of effort. While the mechanism of this fatigue is uncertain, it may relate to the intrinsic brain infection by HIV that also is the cause of AIDS dementia complex. Fatigue is described by HIV patients as “painful” and is one of the most debilitating symptoms that limit their quality of life. In addition to fatigue, sleep disturbances also affect patients with HIV/AIDS, and likely contribute to the severity of their daytime fatigue. Sleep dysfunction, and in particular insomnia, are associated with CNS infection by HIV many years before the onset of AIDS.

The diagnosis of ADC is based on neurological history, testing, and examination and is only diagnosed when the patient's day-to-day activities have already become severely degraded. The need for an objective and early method for detecting neurologic impairment due to HIV is vital for the proper treatment and management of HIV/AIDS. We propose the use of actigraphy analyzed by functional data analysis methods as an objective and early method for detecting neurologic impairment due to HIV

Order-Constrained Solutions in K-Means Clustering: Even Better Than Being Globally Optimal

Douglas Steinley, University of Missouri

Lawrence Hubert, University of Illinois

Abs: An order-constrained K-means cluster analysis strategy is proposed and implemented through an auxiliary quadratic assignment optimization heuristic that identifies an initial object order. A subsequent dynamic programming recursion is applied to optimally subdivide the object set subject to the order constraint. It is shown that although the usual K-means sum-of-squared-error criterion is not guaranteed to be minimal, a true underlying cluster structure may be more accurately recovered. Also, substantive interpretability seems generally improved when constrained solutions are considered. We illustrate the procedure with several data sets from the literature.

A Scalable Approach to Clustering Massive Audio Catalogs

David K. Tcheng, NCSA/ALG, One Llama Media Inc.

Abs: Our goal is to create a general purpose audio clustering algorithm that scales well to the largest of audio data sets. Currently we are processing a 600 hour audio catalog of continuous (24/7) recordings from wireless microphones mounted on birds (territorial cardinals). In addition audio recordings, we have corresponding recordings of the birds X/Y position over time as estimated by radio location. Our first task is to discover the syllables and grammar that fully describe the cardinal utterances. Next we seek to understand bird behavior – how the utterances of a specific bird is influenced by the position of the bird, time of day, season, weather, and

the sounds of other birds and animals in proximity. Our approach begins by creating a continuous spectrogram of the audio waves using a bank of band pass filters. Next we segment the spectrogram using a sliding window of fixed length (e.g., one second). A hash table based clustering algorithm is used to find frequently occurring patterns. A simulated annealing based clustering algorithm is used to arrange the audio spectra segments into a 2-d map which maximizes the relationship between map distance and audio spectra similarity.

Author name disambiguation in MEDLINE: results from first-pass clustering

Vetle I. Torvik and Neil R. Smalheiser

University of Illinois at Chicago

Abs: We have previously described (Torvik et al., JASIST, 2005) a method for automatically generating training sets and estimating the probability that a pair of papers in MEDLINE sharing a last and first name initial are authored by the same individual. The probability estimates were based on shared title words, journal name, co-author names, medical subject headings, language, and affiliation, as well as distinctive features of the name itself (i.e., presence of middle initial, suffix, and prevalence in MEDLINE). This project, which we call “Author-ity”, was subsequently funded by an NIH grant to create a database of all papers in MEDLINE clustered by predicted author-individuals. We have recently completed a first-pass clustering of the 2006 baseline version of MEDLINE (¿ 15 million papers) resulting in ¿ 5 million predicted author-individuals. In this talk, we will summarize our work to date: a) the basic pairwise model with recently added and modified predictive features (first name variants, email address, last name specific affiliation word stoplists); b) new automatic methods of generating large training sets; c) methods for estimating the prior probability for given a population of papers to be compared; d) a weighted least squares algorithm for correcting triplet violations of the form $p_{ij} + p_{ik} \geq 1 + p_{jk}$, e) a database with predicted author-individuals resulting from simple agglomerative clustering; and f) some preliminary data for future research directions.

Robust Partial Logistic Regression (RoPLR)

Asuman S. Turkmen and Nedret Billor

Department of Mathematics and Statistics, Auburn University

Abs: Partial least squares (PLS) is a statistical technique to summarize high dimensional and correlated predictor variables into low dimensional, uncorrelated variables which have the best predictive power. High dimensionality and collinearity make the application of the available classification methods difficult and even impossible for some cases. An analyst can reduce the dimension by constructing new components employing PLS technique and then apply a classification method on the constructed PLS components. Even though PLS was originally designed for continuous response variables, only for last six years, it has become a widely used statistical dimension reduction technique for classification. The classical PLS method is known to be very sensitive to outlying observations that usually exist in experimental data. Therefore several robust PLS methods have been proposed when the response variable is continuous. However, to our knowledge, there has been no study on robustness of PLS methods for dimension reduction in classification. In this study, the effect of outliers on existing PLS classification methods is investigated. We also propose a new PLS algorithm based on robust logistic regression (RoPLR) for classification problems. Real and simulated data sets are used to demonstrate the performance of RoPLR method.

Classification of Self-Modeling Regressions

Rhonda VanDyke, Kert Viele, and Robin Cooper

Departments of Statistics and Biology, University of Kentucky

Abs: A set of self-modeling functions is defined by the entire set of functions being related through affine transformations of the x and y axes to a common function $g(t)$. We expand this definition to include the possibility that a set of functions contains two underlying sets of self-modeling functions, the first related through a common shape $g_1(t)$ and the second related through a separate shape function $g_2(t)$. Our goal is to take data consisting of a set of functions, estimate the two underlying shape functions, and to classify each function as belonging to either the first or second group of self-modeling functions. We estimate the underlying shape functions through Bayesian Adaptive Regression Splines (BARS). We illustrate the methodology through Synaptic Transmission data, where the functions measure electrical current across time and the two self-modeling groups of functions are hypothesized to result from different vesicles within synapses

releasing transmitter in qualitatively different manners.

Beta-Distributed Generalized Linear Mixed Models

Jay Verkuilen, Department of Psychology, University of Illinois Urbana-Champaign

Michael Smithson, School of Psychology, The Australian National University.

Abs: Smithson and Verkuilen (2006) proposed the use of beta regression for interval data with known lower and upper bounds. Variables of this form abound in the behavioral sciences and are frequently not well-modeled by Gaussian error distributions. For instance, confidence ratings, judged probabilities, or leverage ratios of firms are all in the unit interval and other bounded interval variables can be rescaled to the unit interval without loss of generality. These variables frequently exhibit marked skew or floor/ceiling effects. Unlike the Gaussian, the beta distribution better accommodates empirical distributions which may be L- or J-shaped, long tailed, or symmetric. Building on our prior work, we present a regression model that assumes, conditional on fixed predictors and random effects, the dependent variable is beta distributed. Estimation by marginal maximum likelihood or Bayesian MCMC works quite well in practice. The model is illustrated with repeated measures data exhibiting between-subject heteroscedasticity taken from an experiment considering the "LakeWobegon effect".

Multilevel Latent Markov Models for Discrete Longitudinal Data

Hsiu-Ting Yu, Department of Psychology, University of Illinois at Urbana-Champaign

Abs: Multilevel longitudinal data are clustered both temporally and spatially. The longitudinal or repeated measures within subjects aspect of the data extends horizontally along a time dimension. The multilevel or hierarchically nested structure extends vertically on a spatial dimension. Since both types of clustered structures can induce dependency into the data, both aspects need to take into account when modeling the data. Many developments for multilevel and longitudinal data have focused on continuous response or outcome variables; however, less attention has been paid to the data with discrete manifest and latent variables. In this talk, I will review the classical latent class models for discrete data and discuss the current methods of the extension on each of the clustering data structures.

Multilevel Latent Markov Models are proposed to unify the two types of dependency due to clustering in a single model. The proposed models are hybrids of random-effects and conditional models, where conditional model is adopted to model the change between two occasions; and random-effects modeling approach is utilized to account for the effects of nested structure. A data set from Educational Longitudinal Study of 2002 is used to illustrate various models for discrete longitudinal data with multilevel data structure.