Classification Society 2012 Annual Meetings

Carnegie Mellon University, Pittsburgh, PA: June 14th-16th, 2012
(registration and all meeting locations are in Baker Hall, Lower Level)

Wednesday, June 13th Welcome Happy Hour, 6-8pm *Bridges Lounge, Holiday Inn*

Thursday, June 14th

7:45-8:30am: Bagels, Pastries, Coffee *Baker Coffee Lounge*

8:30am: Conference Welcome/Opening Remarks

8:30-10:00am: **Multidimensional Scaling, Data Analysis, and Associated Software**
Chair: Beth Ayers, American Institutes for Research

Stephen France, Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee
Analysis of a Generalized Agreement Metric for Comparing Configurations

Cherise R. Chin Fatt, The University of Texas at Dallas, School of Behavioral and Brain Science
DISTATIS: Three-way metric multidimensional scaling and its extensions

Ulas Akkucuk, Bogazici University, Istanbul, Turkey
Applications of Perceptual Mapping Software on Data Originating from Social Sciences

10:00-10:30am: Coffee Break *Baker Coffee Lounge*

10:30-11:30am: **Classification Applications in Criminal Behavior Patterns**
Chair: Beth Ayers, American Institutes for Research

Tim Brennan, Northpoint Institute
Toward a Multi-Axial Classification for a State Prison System: Balancing Cluster Reliability, Case Assignment Accuracy, and User Requirements

Constantinos Kallis, Forensic Psychiatry Research Unit, Queen Mary, University of London, UK
Computer Instrument for Violence (CIV): A new prognostic model for violent reoffending for UK serious offenders

11:30am-1:00pm: Lunch Break

1:00-1:45pm: **Keynote Address**
Chair: Rebecca Nugent, Dept of Statistics, Carnegie Mellon

Douglas Steinley
Associate Professor, Department of Psychological Sciences, University of Missouri, Columbia

Distinguishing Between Continuous and Discrete Latent Spaces:
Implications for Psychological Research

1:45-2:00pm: Coffee Break *Baker Coffee Lounge*

2:00-3:30pm: **Improving the Utility and Tractability of Classification**
Chair: Stan Sclove, University of Illinois at Chicago

Willem Heiser, Faculty of Social and Behavioral Sciences, Leiden University
Clustering of Rankings using Kemeny Distances

Dave Dubin, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign
On Data Identity and Scientific Equivalence

Hans-Friedrich Koehn, University of Illinois, Urbana-Champaign
A Branch-and-Bound Max-Cardinality Algorithm for Exploratory Mokken Scale Analysis

3:30pm-4:00pm: Coffee Break *Baker Coffee Lounge*

4:00-5:30pm: **Updating (Mixture) Model-Based Clustering I**
Chair: Doug Steinley, Univ of Missouri Psychology

Emilie Rausch, Department of Psychological Sciences, University of Missouri
K-Means and Mixture Model Clustering: A Comparison

Raju Vatsavai, Oak Ridge National Laboratory
Revisiting the Problem of Finding the Optimal Number of Clusters

Yuhong Wei, Department of Mathematics & Statistics, University of Guelph
Mixture Model Averaging for Clustering

6 - 7:30pm: Reception (drinks and light snacks) *Lucca Ristorante*
Reception is at: 317 S. Craig St, Pittsburgh, PA 15213

Friday, June 15th

7:45-8:30am: Bagels, Pastries, Coffee *Baker Coffee Lounge*

8:30-10:00am: **Record Linkage and Disambiguation**
Chair: Rebecca Nugent, CMU Statistics

Mauricio Sadinle, Department of Statistics, Carnegie Mellon University
Multiple Record Linkage Using a Generalized Fellegi-Sunter Framework

Rob Hall, Department of Machine Learning, Carnegie Mellon University
Valid Statistical Inference on Automatically Matched Files

Sam Ventura, Department of Statistics, Carnegie Mellon University
Disambiguating USPTO Inventors with Classification Models Trained on Comparisons of Labeled Inventor Records

10:00-10:30am: Coffee Break *Baker Coffee Lounge*

10:30-11:30am: **Classification Society Dissertation Award Winners**
Chair: Rebecca Nugent, CMU Statistics

Theodore Damoulas, Department of Computer Science, Cornell University
Probabilistic Multiple Kernel Learning

Ondrej Vencalek, Faculty of Mathematics and Physics, Charles University, Faculty of Science, Palacky University, Olomouc
Depth-based Classification

11:30am-1:00pm: Lunch Break (CS Board Meeting in Baker 154R)

1:00-1:45pm: **Presidential Address**
Chair: Doug Steinley, Univ of Missouri Psychology

Rebecca Nugent, Department of Statistics, Carnegie Mellon University

A Self-Tuning Diffusion Framework for Use in Document Clustering

1:45-2:00pm: Coffee Break *Baker Coffee Lounge*

2:00-3:00pm: **Memorial Session Celebrating the Life and Work of Bob Sokal and Phipps Arabie**
Chair: Willem Heiser, Faculty of Social and Behavioural Sciences, Leiden University

Willem Heiser, Faculty of Social and Behavioral Sciences, Leiden University

F. James Rohlf, Department of Ecology and Evolution, Stony Brook University

Mel Janowitz, DIMACS, Rutgers University

3:00-3:30pm: Coffee Break *Baker Coffee Lounge*

3:30-5:00pm: **Advances in Classification/Clustering Methodology**
Chair: Doug Steinley, Univ of Missouri, Psychology

Pascal Cuxac, INIST-CNRS

A New Approach for Improving Clustering Quality Based on Connected Maximized Cluster Features

Jean-Charles Lamirel, Synalp team - LORIA

A New Incremental Clustering Approach Based on Cluster Data Feature Maximization

Rebecca Le, Department of Statistics, University of California, Riverside
Alternative Approaches in Multi-Label Neutral Zone Classification Problems

5:00-5:30pm: Classification Society General Meeting

7:00-10:00pm: Classification Society Banquet *Carnegie Museum of Natural History*
Our banquet is in the Foster Overlook; Museum is at 4400 Forbes Avenue

Saturday, June 16th

8:00-9:00am: Bagels, Pastries, Coffee *Baker Coffee Lounge*

9:00-10:30am: **The Use of Indices for Clustering Quality**
Chair: Beth Ayers, American Institutes of Research

Michaela Hoffman, Department of Psychological Sciences, University of Missouri
Comparing Methods for Link Prediction in Networks

Jean-Charles Lamirel, Synalp team - LORIA
An unbiased symbolico-numeric approach for reliable clustering quality evaluation

Ahmed Albatineh, Florida International University
On the Equivalence of Some Indices of Similarity: Implication to Binary Presence/Absence Data

10:30-11:00am: Coffee Break *Baker Coffee Lounge*

11:00-12:30pm: **Applications in the Biological Sciences**
Chair: Beth Ayers, American Institutes for Research

Tanzy Love, Department of Biostatistics and Computational Biology, University of Rochester
Latent Effect Modification found in the Effect of Fish Consumption on IQ

Gabrielle Flynt, Department of Mathematics, Bucknell University
Clustering Trajectories in the Presence of Informative Monotone MIssingness

Sonia Todorova, Department of Statistics, Carnegie Mellon University
Model-based Clustering of non-Poisson, non-homogenous Point Processes Events with Application to Neuroscience

12:30pm-1:30pm: Coffee/Lunch Break (sandwiches provided) *Baker Coffee Lounge*

1:30-3:00pm: **Updating (Mixture) Model-Based Clustering II**
Chair: Rebecca Nugent, CMU Statistics

Brian C. Franczak, Department of Mathematics & Statistics, University of Guelph
The ParSAL Family

Irene Vrbik, Department of Mathematics & Statistics, University of Guelph
MCLUST Models Extended to Mixture of Multivariate Skew-T Distributions

Jeff Andrews, Department of Mathematics & Statistics, University of Guelph
Using Evolutionary Algorithms for Model-Based Clustering

3:00pm: Conference Closing Remarks

ABSTRACTS

*Presenters are the first authors unless otherwise denoted by an **

Session 1: Multidimensional Scaling, Data Analysis, and Associated Software

Analysis of a Generalized Agreement Metric for Comparing Configurations

Stephen France
Sheldon B. Lubar School of Business, University of Wisconsin - Milwaukee

We describe a family of metrics for measuring agreement between sets of solution configurations. These metrics are based upon the Rand index for measuring cluster agreement and in the literature are described as agreement metrics or measures of local continuity. We review and synthesize the current literature. We describe extensions for measuring confidence intervals and investigating statistical hypotheses. We describe a set of software tools developed in MATLAB and R to implement a broad range of agreement metrics. We demonstrate the use of the agreement metrics using examples from marketing, psychology, and the social sciences. We show how the agreement metrics can be used to evaluate dimensionality reduction methods, tune method parameters, and evaluate how solution configurations change over time.

DISTATIS: Three-way metric multidimensional scaling and its extensions

Cherise R. Chin Fatt, Derek Beaton, Herve Abdi
The University of Texas at Dallas, School of Behavioral and Brain Science

Metric multidimensional scaling (MMDS) transforms a single distance matrix into a set of coordinates such that the Euclidean distance matrix derived from these coordinates best approximate the original distance matrix. DISTATIS is a recent generalization of MMDS that can handle multiple distance matrices measured on the same set of observations. DISTATIS can be seen as a combination of MMDS and STATIS which itself, generalizes principal component analysis to handle a set of data tables measured on the same observations.

The goal of DISTATIS is to obtain a common representation - called a compromise - of the distance matrices. To do so, DISTATIS (like MMDS) transforms each distance matrix into a cross-product matrix, and (like STATIS) computes the compromise as an optimal linear combination of the cross-product matrices. The eigen-decomposition of the compromise gives factors scores that can be used to create maps of the observations. The original distance matrices can also be projected onto the compromise to show how each distance matrix interprets the common space. Out-of-sample (a.k.a. supplementary) distance matrices can also be projected onto the compromise. As a by-product of the computation of the compromise, DISTATIS also provides a map showing the similarity between the original distance matrices. Inferences in DISTATIS are performed using cross-validation techniques. In this talk we will present and illustrate DISTATIS and its implementation in a comprehensive R package.

Applications of Perceptual Mapping Software on Data Originating from Social Sciences

Ulas Akkucuk
Bogazici University, Istanbul, Turkey

Perceptual mapping is a useful technique that could be employed by marketing managers and researchers in the field of product and brand management. With this technique researchers try to gain insight about the relative positioning of brands with respect to one another. There are a number of multivariate methods that could be used to achieve this purpose. Some of these procedures have been incorporated into commonly available software packages. R, Matlab and IBM-SPSS have their own modules for performing the perceptual mapping task. Each of these categories has different input-output considerations and algorithmic properties. In this paper, the main motivation is to illustrate the differences and similarities between the different perceptual mapping software applications with special emphasis

on the two main SPSS modules, ALSCAL and PROXSCAL. For the purpose of demonstrating the different algorithms, data have been collected on similarities between automobile brands and attribute ratings on the same brands. In addition, OECD data about patent applications of different countries in various fields have been downloaded. Comparisons of the solutions are performed not only by using the performance measures reported by the programs, but also by using the agreement rate which can capture the degree of similarity of the input to the output.

Session 2: Classification Applications in Criminal Behavior Patterns

Towards a Multi-Axial classification for a State Prison System: Balancing cluster reliability, case assignment accuracy, and user requirements

Tim Brennan, William L. Oliver
Northpoint Institute

This applied project describes an attempt to build a multi-axial classification for a State Prison System using a large sample of detainees. The separate axes included: 1) Criminal behavior patterns, 2) Causal-Explanatory patterns, 3) Internal Classifications for management and treatment of prisoners, and 4) Predictive classifications at release. We will present the construction of Axes 1,2, and 3. Cluster identification used several standard Bagged K-means approaches, incorporating measures of cluster stability based on resampling. In this on-going project several classification challenges were encountered including: variable selection and weighting during clustering, error estimation and confidence levels for case assignment using random forest (RF) methods. Current results will be presented.

Computer Instrument for Violence (CIV): A new prognostic model for violent reoffending for UK serious offenders

Constantinos Kallis
Forensic Psychiatry Research Unit, Queen Mary, University of London, UK

Background: A number of prognostic methods are currently implemented for serious offenders prior to their release from prison to predict their likelihood of being convicted for violence after release. These methods have been constructed with a variety of statistical approaches and the predictors are selected mainly based on the associations between candidate predictors and the outcome. The predictive accuracy of these methods in terms of the Area under the ROC curve (AUC) is moderate for the UK population of serious offenders but remains unknown for key personality disorder diagnostic groups within this population.

Objectives:

1. To identify methodological problems related to the construction of existing prognostic methods for violent reoffending.

2. To create a novel prognostic model using a new statistical approach that will improve predictive accuracy for the population of UK serious offenders and important diagnostic subgroups when compared to existing methods

Methods: 1380 Prisoner Cohort Study (PCS) male participants have been assessed for the risk for violent reoffending prior to release from prison with established prognostic methods. To construct our prognostic model (CIV), we used data collected prospectively in PCS and the Police National Computer (PNC) .

Results: The overall AUC for CIV is 0.76 (95% C.I. 0.74, 0.79, $p < 0.001$) and it is higher and significantly different when compared to the corresponding AUCs for the other prognostic methods. Using the Hosmer-Lemeshow statistic, we find that the predicted probabilities are well-calibrated. For cases with primary psychiatric (Axis I) disorders, the AUC is 0.75 (95% C.I. 0.71, 0.78, $p < 0.001$) and for those with Antisocial Personality Disorder (ASPD), the corresponding AUC is 0.72 (95% C.I. 0.68, 0.75, $p < 0.001$). These results indicate that the predictive accuracy is relatively similar.

Conclusion: By constructing a prognostic model for violent reoffending that uses only highly predictive items for a specific outcome, it is possible to improve significantly the predictive accuracy of prognostic methods.

Keynote Address

Distinguishing Between Continuous and Discrete Latent Spaces: Implications for Psychological Research

Douglas Steinley
Department of Psychology, University of Missouri-Columbia

Given the revision of the DSM-IV to the DSM-V, there is an increasing interest in distinguishing between the nature of latent spaces that underlie psychological constructs. The main emphasis is determining whether these latent spaces are discrete or continuous. This talk discusses several approaches to answering this question. Some preliminary results are presented and a variety theoretical considerations are discussed.

Session 3: Improving the Utility and Tractability of Classification

Clustering of Rankings using Kemeny Distances

Willem Heiser, *Faculty of Social and Behavioral Sciences, Leiden University, The Netherlands*
Antonio D' Ambrosio, *Dept of Mathematics and Statistics, University of Naples Federico II, Italy*

Rankings and partial rankings are ubiquitous in data analysis, yet there is relatively little work in the classification community that uses the typical properties of rankings. The natural metric for rank order data is the Kemeny distance, defined as the minimum number of interchanges of two adjacent elements required to transform one (partial) ranking into another. The Kemeny distance is equivalent to Kendall's τ for complete rankings, but for partial rankings it is equivalent to Emond and Mason's extension of τ. As a clustering algorithm, we use the flexible class of methods proposed by Ben-Israel and Iyigun, which generalizes the K-means algorithm, and define the disparity between a ranking and the center of a cluster as the Kemeny distance. The median ranking characterizes the location of a cluster, and average extended τ, the homogeneity of a cluster.

On Data Identity and Scientific Equivalence

David Dubin
Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign

Carroll and Arabie's 1980 taxonomy of measurement data and models departs from the data theory of Coombs (1964) in attempting "separate taxonomies of data and models," rather than classifying data with respect to dominance or proximity relations among points in an abstract space. This independence or interdependence bears directly on the problem of defining scientific equivalence for current data-intensive research, and to the complexities of practical issues such as justifying appropriate statistics on the basis of measurement scale type.

A Branch-and-Bound Max-Cardinality Algorithm for Exploratory Mokken Scale Analysis

Michael J. Brusco, *Florida State University, Tallahassee*
*Hans-Friedrich Koehn, *University of Illinois, Champaign Urbana*
Douglas Steinley, *University of Missouri-Columbia*

Exploratory Mokken scale analysis can be conceptualized as a combinatorial optimization problem: from a set of candidate items, a maximal subset must be selected such that (1) the (normed) pairwise item covariances, H_{jk} , are all strictly positive; (2) the item scalability coefficients, H_j , of all items selected exceed a predetermined threshold, c; (3) the set of selected items maximizes the scale coefficient, H . Mokken proposed a stepwise, bottom-up algorithm, relying on a greedy search strategy, termed Automated Item Selection Procedure (AISP), that has been implemented in the commercially distributed software package MSP (the statistical software package Stata also contains a module for Mokken analysis). Recently, AISP has been made freely available in the R package mokken that, as a new development, also offers a genetic algorithm for exploratory Mokken scale analysis.

Among the class of object selection problems, maximum cardinality subset selection requires finding the largest possible subset of objects that satisfies one or more constraints. We present an exact branch-and-bound algorithm for maximum cardinality subset selection tailored to exploratory Mokken scale analysis of a set of binary items. Computational results are reported for simulated data with max 80 items generated from (1) the DINA model; (2) the five-parameter acceleration model; (3) the Rasch model (using different item discrimination parameter settings for generating data sets that mix item subsets all satisfying the double monotonicity condition, while the entire data sets do not) - thus, these data sets represent incrementing challenges to the proposed algorithm.

Session 4: Updating (Mixture) Model-Based Clustering I

K-Means and Mixture Model Clustering: A Comparison

Emilie Rausch
Department of Psychological Sciences, University of Missouri

The clustering of objects to explain an underlying group structure is popular in many scientific fields. One can analyze the cluster structure of their data in one of two ways: 1) using traditional optimization based clustering approaches, or 2) using a parametric approach such as finite mixture modeling. Equivalencies between these two forms of clustering are discussed, and results from a preliminary simulation are given that compare their performance in a variety of circumstances.

Revisiting the Problem of Finding Optimal Number of Clusters

Ranga Raju Vatsavai
Oak Ridge National Laboratory, Tennessee

Finding optimum number of clusters is an age-old problem. We revisit this problem by looking at recent solutions, specifically G-Means and X-Means algorithms. In this paper we present a computationally efficient model-based split and merge clustering algorithm that incrementally finds model parameters and the number of clusters (or components) in a Gaussian Mixture Model (GMM). The basic algorithm we present is similar to that of G-means and X-means algorithms; however, our proposed approach avoids certain limitations of these well-known clustering algorithms that are pertinent when dealing with geospatial data. Additionally, we attempt to provide insights into finding the optimal number of clusters and other data mining challenges that are encountered when clustering geospatial data. We compare the performance of our approach with the G-means and X-means algorithms. Experimental evaluation on simulated data and on multispectral and hyperspectral remotely sensed image data demonstrates the effectiveness of our algorithm.

Mixture Model Averaging for Clustering

Yuhong Wei, Paul McNicholas
Department of Mathematics & Statistics, University of Guelph

Model-based clustering is based on a finite mixture of distributions, where each mixture component corresponds to a different group, cluster, subpopulation, or part thereof. Gaussian mixture distributions are most often used. Criteria commonly used in choosing the number of components in a finite mixture model include the Akaike information criterion, Bayesian information criterion, and the integrated completed likelihood. The best model is taken to be the one with highest (or lowest) value of a given criterion. This approach is not reasonable because it is practically impossible to decide what to do when the difference between the best values of two models under such a criterion is small. Furthermore, it is not clear how such values should be calibrated in different situations with respect to sample size and random variables in the model, nor does it take into account the magnitude of the likelihood. It is, therefore, worthwhile considering a model-averaging approach. We consider an averaging of the top M mixture models and consider applications in clustering and classification. In the course of model averaging, the top M

models often have different numbers of mixture components. Therefore, we propose a method of merging Gaussian mixture components in order to get the same number of clusters for the top M models. The idea is to list all the combinations of components for merging, and then choose the combination corresponding to the biggest adjusted Rand index (ARI) with the reference model. A weight is defined to quantify the importance of each model. The effectiveness of mixture model averaging for clustering is proved by simulated data and real data under the pgmm package, where the ARI from mixture model averaging for clustering are greater than the one of corresponding best model. The attractive feature of mixture model averaging is its computationally efficiency; it only uses the conditional membership probabilities. Herein, Gaussian mixture models are used but the approach could be applied effectively without modification to other mixture models.

Session 5: Record Linkage and Disambiguation

Multiple Record Linkage Using a Generalized Fellegi-Sunter Framework

Mauricio Sadinle, *Dept of Statistics, Carnegie Mellon University*
Stephen E. Fienberg, *Maurice Falk University Professor of Statistics and Social Sciences, Department of Statistics, Carnegie Mellon University*

We present a probabilistic method for linking multiple datafiles. This task is not trivial in the absence of unique identifiers for the individuals recorded. This is a common scenario when linking census data to coverage measurement surveys for census coverage evaluation, and in general when multiple record-systems need to be integrated for posterior analysis. Our method generalizes the Fellegi-Sunter theory for linking records from two datafiles and its modern implementations. The multiple record linkage goal is to classify the record K-tuples coming from K datafiles according to the different matching patterns. Our method incorporates agreement transitivity in the computation of the data used to model matching probabilities (via a mixture model). We present a method to decide the record K-tuples membership to the subsets of matching patterns and we prove its optimality. We apply our method to the integration of three Colombian homicide record systems and perform a simulation study in order to explore the performance of the method under measurement error and different scenarios. The proposed method works well and opens some directions for future research.

Valid Statistical Inference on Automatically Matched Files

Rob Hall, *Department of Machine Learning, Carnegie Mellon University*
Stephen E. Fienberg, *Maurice Falk University Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University*

We develop a statistical process for determining a confidence set for an unknown bipartite matching. It requires only modest assumptions on the nature of the distribution of the data. The confidence set involves a set of linear constraints on the bipartite matching, which permits efficient analysis of the matched data, e.g., using linear regression, while maintaining the proper degree of uncertainty about the linkage itself.

Disambiguating USPTO Inventors with Classification Models Trained on Comparisons of Labeled Inventor Records

Samuel L. Ventura, *Department of Statistics, Carnegie Mellon University*
Rebecca Nugent, *Department of Statistics, Carnegie Mellon University*
Erica R.H. Fuchs, *Department of Engineering & Public Policy, Carnegie Mellon University*

The United States Patent and Trademark Office does not assign identification numbers to records of unique inventors in its database of over 8 million patents, making it difficult to study the patenting trends and characteristics of individual inventors. Existing methods for disambiguating inventor records in the USPTO database are flawed, failing to train on comparisons of labeled inventor records and/or utilize statistical record linkage methods. Using a set of 47,125 labeled inventor records, we show that applying supervised learning techniques to inventor disambiguation substantially reduces the percentage of false positive and negative errors in the results. In particular, we analyze

the effectiveness of several different classification models and find, empirically, that Random Forests yields the best balance of both low false positive and negative errors in the disambiguation results. We then propose a variant of Random Forests which conditions on features of the underlying record-pairs for use in inventor disambiguation, and we show that this "Conditional Forest of Random Forests" further improves the disambiguation results.

Session 6: Classification Society Dissertation Award Winners

Probabilistic Multiple Kernel Learning

Theodore Damoulas
Computer Science, Cornell University

When multiple observers exist for an overall decision making process (e.g multinomial classification), then multiple, and uncertain, sources of information need to be integrated for an overall predictive task. In this talk I will describe recent work, within the probabilistic framework of Bayesian inference, towards this direction. I will provide an overview of the contributed multiple kernel learning models, associated (approximate) inference techniques.

Depth-based Classification

Ondrej Vencalek
Faculty of Mathematics and Physics, Charles University, Prague
Faculty of Science, Palacky University, Olomouc

The first part of the contribution is a short introduction to the data depth. Data depth is an important concept of non-parametric approach to multivariate data analysis. It provides one possible way of ordering the multivariate data. We call this ordering a central-outward ordering. Basically, any function which provides a reasonable central-outward ordering of points in multidimensional space can be considered as a depth function. This vague understanding of the notion of depth function led to the variety of depth functions (such as the halfspace depth, simplicial depth, zonoid depth, L_1-depth), which have been introduced ad hoc since 1970s. The formal definition of a depth function was formulated by Zuo and Sering in 2000 [4].

The second part of the contribution is devoted to the possible applications of the data depth methodology in classification. During the last ten years quite a lot of effort has been put into development of a nonparametric approach, which uses methodology of data depth for solving the classification problem. The idea of using data depth for classification was firstly introduced in paper by Christmann and Rousseeuw in 2001 [1]. Number of researchers followed and broaden the idea of the pioneering paper. The simplest depth-based classifier (so called maximal depth classier) assigns a new observation to the distribution, with respect to which it has maximal depth. The maximal depth classifier is known to be asymptotically optimal (it has the lowest possible average misclassification rate) only under many restrictive assumptions. Thus more advanced (but not so simple) methods are needed.

The last part of the contribution deals with some new results shown in my doctoral thesis [3]. We introduce an idea of modified k-nearest-neighbour classifier. The classifier overcomes some problems of simple depth-based classifiers. The method can be implemented more easily than comparable classifiers based on kernel density estimation. The clue to construction of an effective classifier based on data depth lies in the existence of a relationship between the depth and the density function. Existence of such a relationship can be achived by localization of the halfspace depth via weights, as proposed in [2].

References:

[1] Christmann, A., Rousseeuw, P. Measuring overlap in binary regression. Computational Statistics and Data Analysis, 2001, vol. 37, pp. 65 75.
[2] Hlubinka, D., Kotik, L., Vencalek, O. Weighted data depth. Kybernetika, 2010, vol. 46, no. 1, pp. 125 148.
[3] Vencalek, O. Weighted data depth and depth based discrimination. Doctoral thesis, 2011.
[4] Zuo, Y., Serfling, R. General notion of statistical depth function. Annals of Statistics, 2000, vol. 28, pp. 461 482.

Presidential Address

A Self-Tuning Diffusion Map Framework for Use in Document Clustering

Rebecca Nugent, *Department of Statistics, Carnegie Mellon University*
David Friedenberg, *Battelle Memorial Institute*

Document clustering has been a rich research area, resulting in algorithms for grouping a fixed or streaming corpus when topic labels are unknown or pre-defined. Regardless of approach, most methods suffer from the need to analyze a very high-dimensional space of words in the corpus lexicon. This dimensionality is often reduced prior to analysis via some statistical threshold or common sense heuristic (e.g. removing words like ”the”). It might be beneficial to remove this somewhat subjective decision. Diffusion maps are a powerful tool for identifying complicated structure and reducing dimensionality in a wide variety of applications. Representing the connectivity of a data set, diffusion maps project observations into a space in which standard methods can more easily model the structure. We explore the use of a flexible self-tuning diffusion map framework that incorporates local tuning parameters to capture group structure of varying density, if present, in a corpus of documents. Our work thus far has also shown a decrease in importance of the clustering method choice once in the reduced projected space. Although the primary focus of this talk is the recovery of cluster structure, we also present classification and regression frameworks.

Session 7: Memorial Session Celebrating the Life and Work of Bob Sokal and Phipps Arabie

Robert Reuven Sokal 1926 - 2012

F. James Rohlf
Department of Ecology and Evolution, Stony Brook University

An overview will be given of both the life of Bob Sokal, some of his scientific achievements with special emphasis on his contributions to the field of classification. His role in the founding of the Classification Society and the International Federation of Classification Societies will also be described.

Phipps Arabie

Mel Janowitz
DIMACS, Rutgers University

TBA

Session 8: Advances in Classification/Clustering Methodology

A new approach for improving clustering quality based on connected maximized cluster features

Jean-Charles Lamirel, *Synalp team - LORIA - Nancy - France*
Pascal Cuxac, *INIST - CNRS - Nancy - France*

Wherever usual clustering methods are exploited on high dimensional and sparse datasets, like textual datasets, one frequently occurring problem is their inability to discriminate between feature-independent data subsets. Hence, in such cases, different subsets of such kind might be gathered into the same cluster, as well as, a single subset might be split into different clusters, leading to an important decrease of clustering accuracy.

This paper proposes a new technique combining isolation of connected cluster data subsets based on cluster features maximization and further reconstruction of coherent data subsets (i.e. coherent clusters) based on unsupervised Bayesian reasoning.

Feature maximization is a cluster quality metric which associates each cluster with maximal features i.e., features whose Feature F-measure is maximal. Feature F-measure is the harmonic mean of Feature Recall and Feature

Precision, representing themselves new basic unsupervised cluster quality measures. Unsupervised Bayesian reasoning based on features is an efficient technique of clustering results or models combination that has been recently proposed by Lamirel (2012).

Our experiment is based on two high dimensional textual datasets. The first one is the Reuters-21578 reference collection. The second one is constituted by bibliographical records related to scientific papers covering various research topics. Using two different clustering methods, K-means and SOM, together with both external and endogenous validation labels, we highlight the efficiency of such approach for correcting original clustering results.

References: Lamirel J.-C. (2012), A new diachronic methodology for automatizing the analysis of research topics dynamics: an example of application on optoelectronics research, Scientometrics (To be published).

A new incremental clustering approach based on cluster data feature maximization

Jean-Charles Lamirel
Synalp team - LORIA - Nancy - France

We present in this paper the Incremental Growing Neural Gas with Feature Maximization (IGNGF) clustering method: a new incremental neural winner-take-most clustering method belonging to the family of the free topology neural clustering methods. Like other neural free topology methods such as Neural Gas (NG), Growing Neural Gas (GNG), or Incremental Growing Neural Gas (IGNG), the IGNGF method makes use of Hebbian learning for dynamically structuring the learning space. However, contrary to these methods, the use of a standard distance measure for determining a winner is replaced in IGNGF by feature maximization.

Feature maximization is a new cluster quality metric which associates each cluster with maximal features i.e., features whose Feature F-measure is maximal. Feature F-measure is the harmonic mean of Feature Recall and Feature Precision representing themselves new basic unsupervised cluster quality measures.

The paper details the operating mode of the IGNGF algorithm and illustrates, by the use of both unbiased clustering quality measure and external validation labels, that the method can outperform existing algorithms in the task of clustering of high dimensional heterogeneous data, whilst presenting, for the first time in the domain, a genuine incremental behavior.

Another main advantage of the IGNF algorithm is finally illustrated through its application to the construction of a Verbnet-like classification for French verbs (Falk et al. 2012). It concerns the ability of the algorithm to provide end-users both with a high quality data partition (here, reliable verbs classes) and with highly informative cluster labels (here, significant sub-categorization frames and thematic grids characteristic of each verb class).

References: Falk, I., Gardent C., Lamirel J.-C. (2012), Classifying French Verbs Using French and English Lexical Resources, International Conference on Computational Linguistic (ACL 2012), Jeju Island, Korea, July 2012.

Alternative Approaches in Multi-label Neutral Zone Classification Problems

Rebecca Le, *Department of Statistics, University of California, Riverside*
Daniel Jeske, *Department of Statistics, University of California, Riverside*

Various multi-label classification algorithms are broadly developed in the literature. Each of these existing methods has different strengths and drawbacks, and very few attempts have been performed to address the uncertainty. Recently, neutral zone classifiers have been presented to deal with ambiguous data to improve the accuracy of the estimated classification results by trading off the relatively high penalty of an incorrect classification with a lower penalty for remaining neutral about class membership. The existing neutral zone method was developed to construct multiple single-label classifiers for multi-label data. In this work, we present different classification alternatives for multi-label data using a standard logistic regression model, a generalized linear mixed model and Markov random field to relax the current underlying assumptions. The proposed neutral zone classification methods are implemented and tested on simulation data sets and on the biological data set. Their results suggest that our proposed classification approaches are useful alternatives for practical application when working with multi-label data.

Session 9: The Use of Indices for Clustering Quality

Comparing Methods for Link Prediction in Networks

Michaela Hoffman
Department of Psychological Sciences, University of Missouri

Networks in their many forms are becoming increasingly useful in analyzing data, in various fields. One problem associated with networks is finding the missing edges or links in the data set, related to the problem of predicting future links. Various methods have been devised to solve this problem. They range from the more simplistic measures such as graph distance to more complicated methods like the Katz algorithms. Despite the number and variety of solutions, the success rates of these algorithms leave room for improvement. The goal of this presentation is to provide a preliminary exploration and comparison of agreement indices and their ability to find missing links.

An unbiased symbolico-numeric approach for reliable clustering quality evaluation

Jean-Charles Lamirel
Synalp team - LORIA - Nancy - France

Traditional quality indexes (Inertia, DB, ...) are known to be method-dependent indexes that do not allow to properly estimate the quality of the clustering in several cases, as in that one of complex and highly multidimensional data, like textual data. We thus propose an alternative approach for clustering quality evaluation based on unsupervised measures of Recall, Precision and F-measure exploiting the descriptors of the data associated with the obtained clusters. Two categories of index are proposed, that are Macro-indexes and Micro-indexes. These indexes are generic enough to cover the case of Boolean-valued data descriptions as well as the one of real-valued data descriptions.

Unsupervised Macro-Recall and Macro-Precision, enables to evaluate to what extent a numerical clustering model can be assimilated to a Galois lattice natural classifier. They have opposite behaviors according to the number of clusters. Thus, these indexes permit to estimate an optimal number of clusters for a given method and a given dataset. On their own side, unsupervised Micro-Precision and Micro-Recall indexes makes it possible to evaluate the overall quality of a clustering, result whilst clearly distinguishing between homogeneous and heterogeneous, or degenerated results.

An experimental comparison of the behavior of the classical indexes with our new approach is performed, ranking for low dimensional standard UCI datasets to highly multidimensional textual datasets.

On the Equivalence of Some Indices of Similarity: Implication to Binary Presence/Absence Data

Ahmed Albatineh
Florida International University

Cohen's kappa, a special case of the weighted kappa, is a chance-corrected index used extensively to quantify inter-rater agreement in validation and reliability studies. In this paper, it is shown that in terms of inter-rater agreement for 2x2 tables, for two raters having the same number of opposite ratings, the indices of weighted kappa, Cohen's kappa, Peirce, Yule, Maxwell and Pilliner, and Fleiss are identical. This implies that the weights in the weighted kappa become less important under such assumptions. Equivalently, it is shown that for two partitions of the same data set, resulting from two clustering algorithms, having the same number of clusters with equal cluster sizes, these similarity indices are identical. Hence, an important characterization relating equal number of clusters with same cluster sizes to presence/absence of a trait in a reliability study is formulated. Two numerical examples that exemplify the implication of this relationship are presented.

Session 10: Applications in the Biological Sciences

Latent Effect Modification found in the Effect of Fish Consumption on IQ

Tanzy Love
Department of Biostatistics and Computational Biology, University of Rochester

The Seychelles Child Development Study (SCDS) is examining associations between prenatal exposure to low doses of methylmercury (MeHg) from maternal fish consumption and children's developmental outcomes. Secondary analysis from this study found significant interactions between MeHg and both caregiver IQ and income on 19 month IQ (Davidson et al. 1999). These results are dependent on the categories chosen for these covariates and are difficult to interpret collectively. We estimate effect modification of the association between prenatal MeHg exposure and 19 month IQ using a general formulation of mixture regression. Our model creates a latent categorical group membership variable which interacts with MeHg in predicting the outcome. Group membership and the regression coefficients are estimated simultaneously. We also fit the same outcome model when in addition the latent variable is assumed to be a parametric function of three distinct socioeconomic measures. The results show that children of low income mothers with high IQ and a stimulating home environment have a different response to prenatal MeHg exposure than other children.

Clustering Trajectories in the Presence of Informative Monotone Missingness

Gabrielle Flynt
Department of Mathematics, Bucknell University

Longitudinal studies are a research design in the biomedical, biobehavioral and social sciences where the data are comprised of measurements collected at multiple time points for a group of subjects. These studies are important, because characterizing changes over time is more accurate when the observations are made on the same subjects. An unfortunate obstacle in working with repeated measurements is missing data. Missing data in longitudinal studies can come from subject attrition, dropping out of the study, or missing measurements at different time points throughout the study. Missingness is most often dependent on some unobserved variables and may cause biased estimates in statistical procedures that do not account for the informative missing values.

There are many applications in which we might be interested in characterizing the types of trajectories seen in a population, for example, patients' patterns of recovery or patterns of student learning over time. In practice, we often observe a mixture of trajectory shapes and patterns in a population, rather than just one common trajectory. This mixture can be thought of as separate classes of similar trajectories from within the overall heterogeneous population. Identifying both the number of latent classes as well as their shapes and patterns can provide valuable information about how a population changes over time.

The goal of this work is to analyze longitudinal data by accurately clustering the data into latent classes while taking into account the possibility of informative missingness. In classifying subject trajectories, missing data could lead to misclassification if the missingness is not understood and accounted for in the model. It is probable that different patterns of missingness in individual trajectories will not only have different causes, but will have different effects on subject outcomes. The proposed method combines growth mixture models for trajectory classification with pattern mixture models to account for informative missing data. Results will be shown for a simulated data set as well as a data set that measures clinical depression. Additionally, a new agreement index for comparing two partitions of data will be presented.

Model-based clustering of non-Poisson, non-homogenous point process events with application to neuroscience

Sonia Todorova, *Department of Statistics, Carnegie Mellon University*
Valerie Ventura, *Department of Statistics, The Center for the Neural Basis of Cognition, Carnegie Mellon University*
Steven Chase, *Department of Biomedical Engineering, The Center for the Neural Basis of Cognition, Carnegie Mellon University*

This work is inspired by the spike sorting problem in neuroscience. Spike sorting is the task of clustering electrical pulses (spikes) recorded on an electrode in order to recover the signals of individual neurons. This is typically done by clustering the continuous voltage traces of the spikes. The point process of neuron spike times is then used to decode behavior in brain-computer interfaces. We use a mixture model with mixing proportions varying with the behavioral covariates (Ventura, 2009). This novel approach to spike sorting allows us to use simpler low-dimensional features and to improve the accuracy of decoding from real data.

Session 11: Updating (Mixture) Model-Based Clustering II

The ParSAL Family

Brian C. Franczak, Ryan P. Browne, Paul McNicholas
Department of Mathematics & Statistics, University of Guelph

A family of shifted asymmetric Laplace (SAL) distributions, the ParSAL family, is introduced and used for model-based clustering and classication. These models arise through an eigen-decomposition of the component covariance matrices. A variant of the EM algorithm, that presents a novel technique for dealing with the issue of "infinite" likelihood is developed for parameter estimation by exploiting the relationship with the general inverse Gaussian distribution. The ParSAL family is applied to both simulated and real data to illustrate clustering and classification applications. In these analyses, our family of mixture models are compared to the popular Gaussian approaches. This work concludes with discussion and suggestions for future work.

MCLUST Models Extended to Mixture of Multivariate Skew-T Distributions

Irene Vrbik, Paul McNicholas
Department of Mathematics & Statistics, University of Guelph

With the advancement of computer technology, mixture model-based approaches to clustering have become increasingly popular. In recent work, a robust, flexible mixture modelling approach using the skew-t distribution has been explored. We propose a skew-t analogue of the popular MCLUST models that impose various constraints on the eigenvalue decomposed covariance matrices. An exact EM algorithm is outlined and our approach is applied to some benchmark clustering datasets.

Using Evolutionary Algorithms for Model-Based Clustering

Jeff Andrews, Paul McNicholas
Department of Mathematics & Statistics, University of Guelph

In mixture model-based clustering, parameter estimation is generally carried out using the expectation-maximization (EM) algorithm, or some closely related variant. We present a new approach by casting the model-fitting problem as a single-objective evolutionary algorithm. We focus mutations on the component indicator variables and present the rationale for this particular form of mutation. The appeal of an evolutionary algorithm is its ability to more thoroughly search the parameter space, providing an approach inherently more robust with respect to local maxima. This approach is illustrated through application on clustering data sets where comparisons are drawn with traditional model-fitting algorithms.