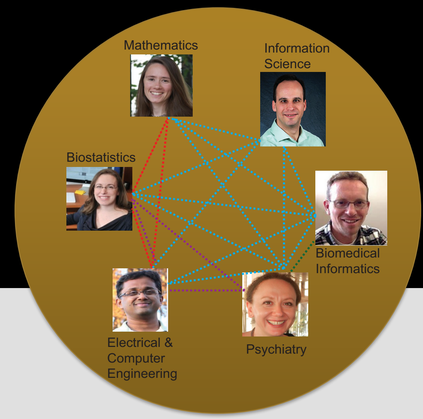
Flexible bootstrapping and analytic approaches towards the clustering of complex medical data

Rachael Hageman Blair[1], Brian Chapman[2], Arianna Di Florio[3,7], Ellen Eischen[4], David Gotz[5], Mathews Jacob[6] and Han Yu[1]

[1]Department of Biostatistics, University of Buffalo; [2]Department of Biomedical Informatics Research and Department of Radiology, University of Utah; [3]Department of Psychiatry, University of North Carolina School of Medicine; [4]Department of Mathematics, University of Oregon; [5]School of Information and Library Science, University of North Carolina; [6]Department of Electrical and Computer Engineering, University of Iowa, [7]Institute of Psychological Medicine and Clinical Nueroscience



Abstract

Identifying subgroups from a severely heterogeneous population is major challenge for Big Data. Different clustering methods optimize differently and consequently capture different aspects of relatedness in the population. Since there is not a one size fits all solution, and no gold standard, the selection of a clustering method can be daunting and problematic. Our interdisciplinary team is working towards the development of interactive ensemble methods for clustering Big Data.

In this first year, we have begun to lay the methodological foundation through the development of a non-parametric bootstrapping approach to estimate the stability of a clustering method. We have developed two novel approaches to bootstrapping stability, and accompanying visualizations, that accommodate different model assumptions, which can be motivated by an investigator's trust (or lack thereof) in the original data. Our approaches outperform state of the art methods for simulation and real data sets of moderate size.

A long term vision of our work is to extend this bootstrapping approach to improve classification and diagnosis of mood disorders, in particular bipolar disorder and major depressive disorder, using data from the UK Biobank. This endeavor would require automated feature selection, sophisticated visualizations, and methods that accommodate mixed data, while retaining valuable clinical interpretations. This project is motivated by the hypothesis that a more precise and personalized classification of mental health disease can be obtained through the development of novel clustering methods that identify clinically significant structures with large population data sets.

One Year Results

Bootstrapping Stability (Yu et al.)
Development of two novel bootstrapping approaches to estimate stability in clustering methods.

- Can be used to estimate stability at three levels:
  1) The stability of the overall method used for clustering.
  2) The stability of individual clusters
  3) The stability of an observation to a given cluster.
- Can be used to determine the optimal number of clusters.
- Performance often superior to state of the art techniques.
- Visualizations facilitate multi-level interpretations of stability.

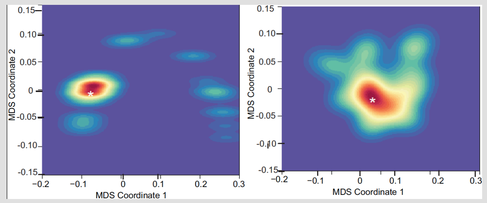


Figure 1: Multi-Dimensional Scaling (MDS) density plots are constructed according to the Jaccard index-based distance between re-sampled cluster labels for (left) iris and (right) wine data

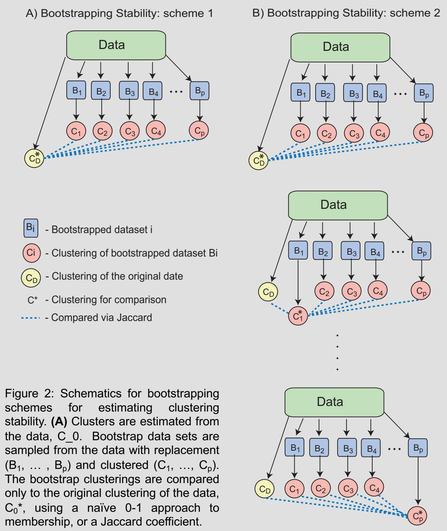


Figure 2: Schematics for bootstrapping schemes for estimating clustering stability. **(A)** Clusters are estimated from the data, C_0. Bootstrap data sets are sampled from the data with replacement (B_1, … , B_p) and clustered (C_1, …, C_p). The bootstrap clusterings are compared only to the original clustering of the data, C_0*, using a naïve 0-1 approach to membership, or a Jaccard coefficient.

(B) Similar to scheme A, clusters are estimated from the data and bootstrapped datasets. However, in addition to comparing the original data clustering to the bootstrapped clusterings, each of the bootstrapped clusterings is compared with each other, and the original data clustering.

Table 1: Performance for identifying the number of clusters, k, for six different simulations of 50 data sets each. Results are shown for prediction strength (pred str), bootstrapping propsed by Fang et al. (Boot2012), and bootstrapping scheme 1(Boot-min-S1) and 2 (Boot-min-S2). The asterisk * indicates true number of clusters.

	Estimated number of clusters						
Method	1	2	3	4	5	6	≥7
Null Model							
Pred Str	46*	4	0	0	0	0	0
Boot2012	0*	4	0	0	0	0	46
Boot-min-S1	47*	3	0	0	0	0	0
Boot-min-S2	46*	4	0	0	0	0	0
Three-cluster model							
Pred Str	0	0	50*	0	0	0	0
Boot2012	0	12	38*	0	0	0	0
Boot-min-S1	0	0	50*	0	0	0	0
Boot-min-S2	0	0	50*	0	0	0	0
Random four-cluster in three dimensions							
Pred Str	0	0	0	50*	0	0	0
Boot2012	0	5	7	38*	0	0	0
Boot-min-S1	0	1	2	47*	0	0	0
Boot-min-S2	0	1	1	48*	0	0	0
Random ten-cluster in three dimensions							
Pred Str	2	3	7	38*	0	0	0
Boot2012	0	13	11	26*	0	0	0
Boot-min-S1	3	5	7	35*	0	0	0
Boot-min-S2	3	3	7	37*	0	0	0
Two elongated clusters							
Pred Str	0	46*	0	4	0	0	0
Boot2012	0	50*	0	0	0	0	0
Boot-min-S1	0	47*	0	3	0	0	0
Boot-min-S2	0	48*	0	2	0	0	0
Two close elongated clusters							
Pred Str	2	35*	12	1	0	0	0
Boot2012	0	34*	6	2	4	0	4
Boot-min-S1	5	40*	4	1	0	0	0
Boot-min-S2	5	41*	3	1	0	0	0

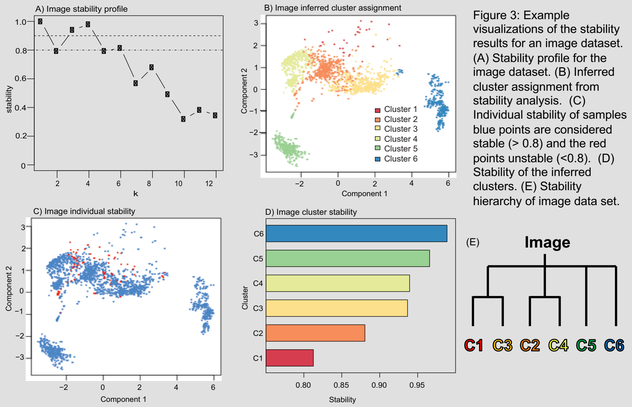


Figure 3: Example visualizations of the stability results for an image dataset. (A) Stability profile for the image dataset. (B) Inferred cluster assignment from stability analysis. (C) Individual stability of samples blue points are considered stable (> 0.8) and the red points unstable (<0.8). (D) Stability of the inferred clusters. (E) Stability hierarchy of image data set.

Convex Clustering of Partially Observed Data (Poddar et al.)
Development of a convex optimization problem to jointly perform clustering and recovery of data with missing entries.

- Relies on partial distance between observations.
- Used for MR image reconstruction, where images are clustered by cardiac phases.
- Instead of a single solution, a cluster-path of solutions is obtained.
- Future applications to mood disorder data.

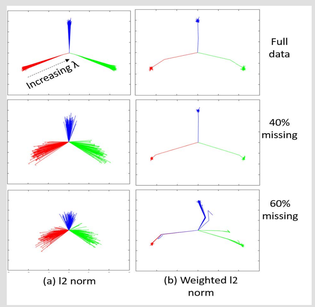


Figure 4: Simulation result to visualize the cluster path for varying degrees of missing data. Cluster paths rely on partially observed data and start to naturally degrade only in severe cases (e.g., 60% missing data).

Five Year Goals

Focus on Mental Health
Three landmark naturalistic studies funded by the National Institute of Mental Health (NIMH) provided some sobering statistics in this respect: psychiatric interventions are **effective in less than 25% of patients** presenting with an acute episode. Diagnoses of mental health conditions are currently characterized by the following:

- Based on little objective evidence (almost arbitrary)
- No biological markers
- Co-morbidity (one person having multiple diagnoses)
- Heterogeneity within diagnosis (two patients with the same diagnosis can have two different sets of symptoms, with little or no clinical overlap)

Data: UK Biobank

- 500,000 patients
- Demographics, survey data, genetic data, and clinical measurements.

Precision Medicine

- Our objective new algorithms and visual tools for precision classification and diagnosis of patients with mood disorders.
- The rigorous identification of subgroups of individuals within heterogeneous populations will facilitate accurate and targeted diagnosis, and provide opportunity for personalized evidence-based interventions.

Data Science Methodology
Our proposed methodology will have the following components:

1. A weighted ensemble based on bootstrapped stability that combines across different clustering methods.

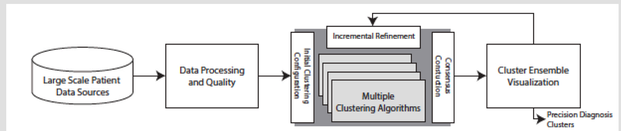


Figure 5: A schematic of the proposed methodology.

1. Efficient interactive clustering algorithms that can handle missing and mixed data-types, as well as severe heterogeneity and feature selection.
2. Interactive visualizations of clusters, which will ultimately aid clinicians.

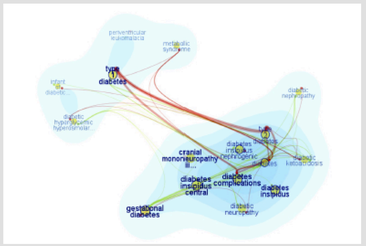


Figure 6: The FacetAtlas visualization (Cao et al.) uses multi-dimensional scaling (MDS) to position data items within a 2D visualization canvas based on pairwise similarity. Kernel density estimation (KDE) is used to compute the blue contours to summarize the dataset, while prominent nodes are rendered individually as landmarks. This view shows diagnosis data with dominant clusters corresponding to Type 1 and Type 2 diabetes.

Broader Impacts

- The proposed approach promises to enable more accurate and targeted diagnosis, and provide personalized evidence for treatment. Given the widespread impact of mental disorders, these techniques have the potential to significantly improve health outcomes for millions of patients.
- The methodology is generalizable to other areas of medicine, beyond mental health, where similar diagnosis and treatment challenges are faced.

References

- Cao N, Sun J, Yu-Ru L, Gotz D, Sixia L. and Huamin Qu. "Facetatlas: Multifaceted visualization for rich text corpora." *IEEE transactions on visualization and computer graphics* 16.6 (2010): 1172-1181.
- Craddock, N., and Mynors-Wallis, L. Psychiatric diagnosis: Impersonal, imperfect and important. British Journal of Psychiatry 204, 2 (2014), 93-95.
- Poddar S and Jacob M “Recovery of partially observed data appearing in clusters”. *Proceedings of the International Conference on Image Processing*. (in press)
- Yu H, Chapman B, DiFlorio A, Eischen E,ßßß Gotz D, and Blair RH. “Bootstrapping estimates of stability for clusters, observations and model selection”. (submitted)

This project is supported by NSF Grants DMS-1557642, DMS-1557593 ,DMS-1557589, DMS-1557576 and DMS-1557668.