Interactive Ensemble Clustering for Mixed Data with Application to Mood Disorders

Rachael Blair Hageman, PhD1; Brian Chapman, PhD2; Arianna Di Florio, MD, PhD3; Ellen Eischen, PhD4; David Gotz, PhD5; Mathews Jacob, PhD6

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

Introduction

Mental disorders are among the most elusive conditions in medicine and defy simple models, be they biological, psychological, social, or any simplistic adjectives. [CMW14] in contrast to current classifications in other areas of medicine, those used in psychiatry, including the Diagnostic and Statistical Manual of Mental Disorders (DSM) [Aus13], rely on clinical manifestations (signs) and subjective reports (symptoms) rather than on the underlying causes and mechanisms.

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.17, 27 Diagnoses of mental health conditions are currently characterized by the following:

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

In this project, we aim to develop a novel quantitative, big-data approach to enable precision diagnosis and treatment in this challenging application domain, with the ultimate goal providing data-driven tools that help clinicians significantly focus patient diagnosis and improve mental health outcomes.

Implications for Precision Medicine and Psychiatry

This project aims to develop an appropriate big-data analysis pipeline for mental health data so as to inform precision diagnosis and treatment in this difficult application domain. Ultimately, the proposed quantitative analysis methods will provide data-driven tools — include interactive visualizations — that will significantly improve patient diagnosis and mental health outcomes. In particular, the project aims to:

1. Enable more precise, personalized diagnoses
2. Enable more precise, personalized treatment decisions
3. Enable the identification of new treatment strategies

Novel and Innovative Methods

The proposed methodology hypothesizes that a more precise and personalized classification of mental health diseases can be obtained through the development of novel integrated clustering methods that identify clinically significant structures within large population datasets. Such methods must:

1. Link treatment response to a very large number of variables, including clinical, virological, biochemical, and psycho-social factors.
2. Overcome several real-world data challenges, including sparsity, ambiguity/uncertainty, missing data, high dimensionality, and heterogeneity.
3. Enable interactively, allowing the user to inject prior knowledge into the classification process.

The key components of our proposed methodology include development of:

1. A weighted ensemble approach that combines self-organizing maps and topological data analysis (including persistent homology)
2. A bootstrapping framework to estimate error
3. An efficient interactive clustering algorithm that can handle missing and mixed data-types
4. Interactive visualizations of clusters, which will ultimately aid clinicians

One-Year Plan

- In the first year, the primary focus is on data heterogeneity, which arguably represents one of the most fundamental statistical concerns.
- This phase focuses on mood disorders (bipolar disorder and major depression) and their genetic and clinical heterogeneity as a simpler problem to develop and test the proposed methodologies. Reasons for this choice include the fact that mood disorders impact over 20% of U.S. adults at some point in life with an economic impact of $210 billion [GFS+15] and that the distinction between unipolar and bipolar depression remains subjective (thus with substantial potential for exploration of objective boundaries, based on the investigations of large data sets).
- Goals for year one include:
  - Identification of data sources
  - Preliminary investigation of algorithms to handle imperfect data, including missing and noisy data
  - Preliminary development of weighted ensemble clustering algorithms, focused on self-organizing maps and persistent homology
  - Static (not interactive) visualization of the clustering results using dimension reduction

Five-Year Goals

Data Science Methodology

The long-term goal is to produce novel algorithms that can address clustering and imputation under the same umbrella, thus being robust to missing and corrupted data. The algorithms will also be capable of accounting for additional priors, which will enable adaptive clustering.

In addition, the project aims to produce novel bootstrapping and consensus algorithms robust to imperfect data to determine the optimal parameters

Visualization Aim

Novel visualization algorithms will facilitate interactive clustering of data, coupled with adaptive clustering.

Precision Medicine Aim

A key goal is to predict treatment response failure (Validation) and suggest potential paths for biologic exploration for alternative therapies

Broader Impacts

Mental Health:

Although the specific focus of the project is mood disorders, the methodologies will be applicable to a broad range of other mental health disorders which suffer from similar challenges of diagnosis and treatment. Therefore, the novel methods outlined in this proposal will lay the foundation for the future development of data-driven, personalized medicine tools for mental health. In particular, 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.

Other Areas of Medicine:

The methodology is generalizable to other areas of medicine, beyond mental health, where similar diagnosis and treatment challenges are faced.

References


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