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

Rachael Hageman Blair¹, Brian Chapman², Arianna Di Florio³,⁷, Ellen Eischen⁴, David Gotz⁵, Mathews Jacob⁶ and Han Yu¹

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; ⁷Institute of Psychological Medicine and Clinical Nueroecence

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 no 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 major depressive disorder and major depressive disorder, using data from the UK Biobank. This endeavor would require automated feature selection, sophisticated visualizations, and methods to accommodate missing 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.

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 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.

2. Interactive visualizations of clusters, which will ultimately aid clinicians.

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 people.

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

References


• Poddar S and Jacob M. “Recovery of partially observed data appearing in clusters”, Proceedings of the International Conference on Image Processing. (in press)