Temporal clustering of biomedical data: starting small, aiming big

Steven Kiddle[1], Elizabeth Baker[1], Maximilian Kerz[1], Caroline Johnston[1,2], Amos Folarin[1,2], Matthew Broadbent[2], Gayan Perrera[3], Rob Stewart[2,4] and Richard Dobson[1,2]

[1] MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, London, UK; [2] NIHR Biomedical Research Centre for Mental Health and Biomedical Research Unit for Dementia at South London and Maudsley NHS Foundation, London, UK; [3] Department of Old Age Psychiatry, IoPPN, King's College London, London, UK. [4] Department of Psychological Medicine, IoPPN, King's College London, London, UK.

Introduction

Increasingly, biomedical researchers have access to big data relating to health, and yet it can be a huge challenge to use this effectively. Common sources of biomedical big data include Electronic Health Records (EHR) and remote monitoring technologies (RMTs; including smartphones and wearables). EHRs and RMTs can provide rich longitudinal data on large numbers of individuals and on many different variables. We are aiming to use this data to better understand disease progression, and to study factors driving disease heterogeneity. To this end, we are developing a clustering approach called Temporal Clustering, which can both cluster and align highly heterogeneously sampled time series. The idea is that this should reveal qualitatively different disease trajectories, which can then be related to treatments, risk factors and health outcomes.

Methods

Clustering consists of a set of steps that we have shown to be effective in a simulation study:

1. Smooth/interpolate time series using smoothing splines
2. Split time series into long and short sets
3. Cluster the long time series using a modified version of k-mean alignment
4. Align the short time series to clusters post-hoc

Results

We are applying the method to: simulated data, FitBit (a RMT) heart rate measures during sleep, and markers of Alzheimer's disease (AD). The AD application separately looks at data from a rich cohort study, and data extracted from a pseudo-anonymised version of the South London and Maudsley (SLaM) NHS Foundation Trust EHR. We will show the results of the simulation study, followed by results from the FitBit and AD examples.

Conclusion

We have developed a clustering methodology that can deal with misaligned time series and highly heterogeneously sampled noisy time series. A fully probabilistic version will be helpful, but we also need to adapt the method to work on big data. It will be necessary to adapt it to handle larger numbers of time series (up to 2 million), and also a larger number of variables.