pHealth 2024
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI240070

Clustering Pseudo Time Series: Exploring Trajectories in the Ageing Process

Puccio Barbara [a,b,1], Tucker Allan [b] and Veltri Pierangelo [c]
[a] *Dept of Surgical and Medical Sciences,University of Catanzaro*
[b] *Department of Computer Science, Brunel University, UK*
[c] *DIMES University of Calabria*

Abstract. Investigating the natural ageing process typically involves the use of extensive longitudinal datasets that can capture changes associated with the progression of ageing. However, they are often resource-intensive and time-consuming to conduct. Cross-sectional data, on the other hand, provides a snapshot of a population at many different ages and can capture many disease processes but do not incorporate the time dimension. Pseudo time series can be reconstructed from cross sectional data, with the aim to explore dynamic processes (such as the ageing process). In this paper we focus on employing pseudo time series analysis on cross-sectional population data that we constrain using age information to create realistic trajectories of people with different degrees of cardiovascular disease. We then use clustering methods to construct and label trajectory-based phenotypes, aiming to enhance our understanding of ageing and disease progression.

Keywords. pseudo time series, clustering analysis, disease progression, ageing

1. Introduction

Investigations into the natural ageing process and its impact on disease development have predominantly depended on gathering and analyzing longitudinal data. These datasets are crucial for monitoring temporal variations in individuals, shedding light on the intricate relationship between ageing and health. Nonetheless, obtaining longitudinal information is beset with obstacles, particularly due to the substantial time and resources required to regularly measure clinical tests on the same individuals. In contrast to longitudinal studies, cross-sectional research provides a snapshot of a population at a single point in time. While this approach is less resource-intensive and offers a broad population level overview of disease prevalence and characteristics, it lacks the temporal dimension necessary to understand how diseases progress or how ageing impacts health over time [1].
Pseudo time series analysis is able to build realistic trajectories through non-time series data based upon appropriate distance metrics and features knowledge. Age information of patients can be used to create more realistic trajectories, thereby enhancing our ability to describe and comprehend the ageing process. By constraining the generation of pseudo time-series based on age information we are able to identify distinct age-related points in trajectories that are associated with specific disease states. The use of clustering allows us to group and label pseudo time-series into different types of

[1] Corresponding Author: Puccio B, barbara.puccio@unicz.it.

trajectory, discerning unique patterns or variations of progression within the dataset, thereby gaining a deeper comprehension of different processes through which ageing and disease manifest.

2. Methods

The idea behind pseudo time series (PTS) is to exploit resampling, distance metrics, and assigned class labels to build realistic trajectories from one label state to another [2]. Let a cross sectional dataset D be defined as a matrix of *m* by *n*, where *m* (rows) is the number of samples and *n* (columns) is the number of clinical features. A full distance matrix is computed through the application of Euclidean distances [3]. Thus, a graph is built and its minimum spanning tree is computed. This step allows to understand the minimal connections that span all data points without creating loops, providing insights into the underlying structure of the data. Subsequently, the shortest path between two points is identified, highlighting the most direct progression between from 'early' states to 'advanced' states.
A PTS is created to show how the trajectories progress from a starting vertex (young state) to an ending vertex (old state). This approach could discover trajectories not only from young to old states, highlighting natural ageing processes but also between different disease states. The PTS points are extracted from the model to perform a clustering analysis to group and label PTS into different types of trajectory and detect unique patterns or variations of progression within the dataset.

3. Results

We applied the method described above to analyze a cross-sectional multivariate dataset of patients diagnosed with heart disease. The dataset is assigned two class labels, that contain age information, where one is deemed to be the starting class or the "young" class and the other as "old" class. The distribution of heart disease status over pseudo time is mapped into 4 values, from 0-healthy to 4-advanced status. Results shows that there is a general trend for younger categories to be correctly identified earlier in pseudo time and later older categories to be identified with later points.

Acknowledgment

B.P. PhD fellows is supported by Relatech S.p.A. and by the Next GenerationEU PNRR DM 352/22. She is visiting at Brunel University. Pierangelo Veltri is supported by SERICS (PE00000014) under the MUR PNRR Next GenerationEU.

References

[1] Tucker A, Li Y, Garway-Heath D, Updating Markov models to integrate cross-sectional and longitudinal studies, Artif Intell Med (2017), doi: https://doi.org/10.1016/j.artmed.2017.09.009
[2] Campbell, K.R, Yau, C. Uncovering pseudotemporal trajectories with covariants from single cell and bulk expression data. Nat Commun 9, 2442 (2018). https://doi.org/10.1038/s4167-018-04696-6
[3] Guzzi P, DiPaola L, Puccio B, Lomoio U, Giuliani A, Veltri P Computational analysis of the sequence structure relation in SARS-CoV-2 spike protein using protein contact networks,Sci Rep (2023)