Artificial Intelligence in Medicine 77 (2017) 23–30

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Updating Markov models to integrate cross-sectional and longitudinal studies

Allan Tucker[a,*], Yuanxi Li[a], David Garway-Heath[b]

[a] Department of Computer Science, Brunel University, UK
[b] Moorfields Eye Hospital and UCL Institute of Ophthalmology, University College London, UK

ARTICLE INFO

Article history:
Received 28 February 2017

Keywords:
Disease progression
Cross-sectional studies
Markov models

ABSTRACT

Clinical trials are typically conducted over a population within a defined time period in order to illuminate certain characteristics of a health issue or disease process. Cross-sectional studies provide a snapshot of these disease processes over a large number of people but do not allow us to model the temporal nature of disease, which is essential for modelling detailed prognostic predictions. Longitudinal studies, on the other hand, are used to explore how these processes develop over time in a number of people but can be expensive and time-consuming, and many studies only cover a relatively small window within the disease process. This paper explores the application of intelligent data analysis techniques for building reliable models of disease progression from both cross-sectional and longitudinal studies. The aim is to learn disease 'trajectories' from cross-sectional data by building realistic trajectories from healthy patients to those with advanced disease. We focus on exploring whether we can 'calibrate' models learnt from these trajectories with real longitudinal data using Baum–Welch re-estimation so that the dynamic parameters reflect the true underlying processes more closely. We use Kullback–Leibler distance and Wilcoxon rank metrics to assess how calibration improves the models to better reflect the underlying dynamics.

1. Introduction

Degenerative diseases such as cancer, Parkinson's disease, and glaucoma are characterised by a continuing deterioration to organs or tissues over time. This monotonic increase in severity of symptoms is not always straightforward however. The rate can vary in a single patient during the course of their disease so that sometimes rapid deterioration is observed and other times the symptoms of the sufferer may stabilise (or even improve – for example when medication is used). Interventions such as medication or surgery can make a huge difference to quality of life and slow the process of disease progression but they rarely change the long term prognosis. The characteristics of many degenerative diseases are therefore a general transition from healthy to early onset to advanced stages. Longitudinal studies [1] measure clinical variables from a number of people over time. Often, the results of multiple tests are recorded, generating Multivariate Time-Series (MTS) data. This is common for patients who have high risk indicators of disease where they are monitored regularly prior to diagnosis. For example, patients with high intra-ocular pressure are brought in to the clinic for visual field tests every six months as they are at high risk of developing glaucoma. The advantages of longitudinal data are that the temporal details of the disease progression can be determined. However, the data is often limited in terms of the cohort size, due to the expensive nature of the studies. Cross-sectional studies record attributes (such as clinical test results and demographics) across a sample of the population, thus providing a snapshot of a particular process but without any measurement of progression of the process over time [2]. An advantage of cross sectional studies is that they capture the diversity of a sample of the population and therefore the degree of variation in the symptoms. The main disadvantage of such studies is that the progression of disease is inherently temporal in nature and the time dimension is not captured. For longitudinal analysis, the patients are usually already identified as being at risk and therefore, controls are usually not available and the early stages of the disease may have been missed. While many data integration techniques address representation heterogeneity where similar data is stored in many different forms [8], as is common in bioinformatics data [26], they do not attempt to combine

* Corresponding author.
E-mail address: allan.tucker@brunel.ac.uk (A. Tucker).

http://dx.doi.org/10.1016/j.artmed.2017.03.005
0933-3657/

variables from cross-section and longitudinal studies, which is what is the focus of this paper. A related area of research, known as panel analysis [21], involves trying to build models along both the temporal dimension and the population dimension from panel studies. Another line of research known as pooling has explored combining cross-sectional data with time-series data [22]. Fitting trends through data [23] is a common approach and is related in some ways to the idea of identifying a trajectory. Another related area of research is sequence reconstruction. This involves trying to find the best order for a particular set of data. Methods include the travelling-salesman-problem approach that aims to minimise the distance between each datum [24], and more recently, the use of PQ trees has been explored to encode partial orderings in order to account for uncertainty in the data due to elements such as noise [25]. Statistical process control [29,28] has also been explored for modelling clinical data including data with unknown temporal ordering. Additionally, a resampling approach known as the Temporal BootStrap (TBS) [5] has been developed that aims to build multiple trajectories through cross sectional data in order to approximate genuine longitudinal data. These 'pseudo time-series' (PTS) can then be used to build approximate temporal models for prediction. This approach has been extended in order to cluster important stages in disease progression using Hidden Markov Models (HMMs) [6]. However, the use of cross-sectional data alone will mean that no genuine timestamps have been used to infer the models and so they only capture an ordering without real temporal information.

In this paper, we explore how to minimise the expensive process of longitudinal data collection by taking models that are learnt from cross-sectional studies using pseudo temporal methods and 'calibrating' with limited longitudinal data. We do this calibration by using the Baum–Welch algorithm to update stochastic models learnt from pseudo time-series so that the dynamic parameters better reflect the underlying process. Essentially, we are integrating cross-sectional and longitudinal data to increase the temporal information and the diversity of data from a large population. Many data integration techniques address representation heterogeneity where similar data is stored in many different forms, as is common in bioinformatics data [7]. Meta Analysis, a popular approach [9], works by supplying a statistical framework for identifying significant results over a number of independent published studies, and calculating the significance of all of the studies when they are brought together. However, it can be prone to publication bias where positive results are more likely to be published and therefore skew the statistics.

In the next section we formally describe the construction of pseudo time-series using the Temporal BootStrap, the experimental set up for assessing the calibration of models with longitudinal data, and the clinical data from glaucoma patients that is used. In the results section, the added value of calibrating pseudo time-series models is demonstrated on simulated data and real clinical data. Finally a case study is explored using the longitudinal glaucoma data and a cross-sectional glaucoma study before conclusions are made.

2. Methods

2.1. Generating pseudo time-series

Let a dataset D be defined as a real valued matrix where m (rows) is the number of samples – here patients – and n (columns) is the number of variables – clinical test data. We define $D(i)$ as the ith row of matrix D. The vector $C = [c_1, c_2, \ldots, c_m]$ represents defined classes, where each $c_i \in \{0, 1\}$ corresponds to the sample i, $c_i = 0$ represents that sample i is a healthy case, and $c_i = 1$ represents that sample i is a diseased case. These classifications are based upon the diagnoses made by experts. We define a time-series as a real valued T (row) by n (column) matrix where each row corresponds to an observation measured over T time points. We say that if $T(i)$ was observed before $T(j)$ then $i < j$.

We define a set of pseudo time-series indices as $P = \{p_1, p_2, \ldots, p_k\}$ where each p_i is a T length vector where $T > 0$. We define p_{ij} as the jth element of p_i and each $p_{ij} \in \{1, \ldots, m\}$. We define the function $F(p_i) = [p_{i1}, \ldots, p_{iT}]$ as creating a T by n matrix where each row of $F(p_i) = D(p_{ij})$. A pseudo time-series can be constructed from each p_i using this operator. For example, if a pseudo time-series index vector $p_1 = [3, 7, 2]$ then $F(p_1)$ is a matrix where the first row is $D(3)$, the second row is $D(7)$ and the third row is $D(2)$. The corresponding class vector of each pseudo time-series generated by $F(p_i)$ is given by $G(p_i) = [C(p_{i1}), \ldots, C(p_{iT})]$.

To demonstrate this notation consider the following example:

Let the data matrix D be defined as:

$$D = \begin{vmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \\ d_{41} & d_{42} & d_{43} \end{vmatrix}, \quad D_{ij} \in \Re.$$

Let the corresponding class vector be $C = [c_1, c_2, c_3, c_4]$. If $P = p_1, p_2$ where $p_1 = [1, 3, 1]$ and $p_2 = [2, 3, 1]$ then:

$$F(p_1) = \begin{vmatrix} d_{11} & d_{12} & d_{13} \\ d_{31} & d_{32} & d_{33} \\ d_{11} & d_{12} & d_{13} \end{vmatrix}, \quad G(p_1) = [c_1, c_3, c_1],$$

and

$$F(p_2) = \begin{vmatrix} d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \\ d_{11} & d_{12} & d_{13} \end{vmatrix}, \quad G(p_2) = [c_2, c_3, c_1].$$

Building pseudo time-series involves plotting trajectories through cross-sectional data based upon distances between each point using prior knowledge of healthy and disease states. These trajectories can then be used to build temporal models such as Dynamic Bayesian Networks (DBNs) [10] and Hidden Markov Models (HMMs) to make forecasts [11]. The Temporal BootStrap involves resampling data [14] from a cross-sectional study and repeatedly building trajectories through the samples in order to build more robust time-series models. Each trajectory begins at a randomly selected datum from a healthy individual and ends at a random datum classified as diseased. The trajectory is determined by the shortest path of Euclidean distances between these two points. The data is first standardised to a mean μ of zero and a standard deviation σ of one as we found that this led to better HMM models. We use the Floyd–Warshall algorithm [12], a well established algorithm used to find the shortest path in a minimum spanning tree from the weighted graph. A full description of the algorithm to generate pseudo time-series is shown in Algorithm 1 and appears in [5]. An example of pseudo time-series that have been generated from cross-sectional data are shown in Fig. 1. Again, this was plotted on the first two components that were generated using multidimensional scaling.