Akmedoids R package for generating directionally-homogeneous clusters of longitudinal data sets

3 Jan 2020

Summary

In many social and behavioural sciences, longitudinal clustering is widely used for identifying groups of individual trends that correspond to certain developmental processes over time. Whilst the popular clustering techniques, such as the k-means and group-based trajectory modelling (GBTM), are more suited for identifying spherical clusters [@GenoFali2010; @Curman2015], their malleability provide the perfect opportunities for identifying other forms of clusters, including those that represent linear growth over time (i.e. the directionally-homogeneous clusters). We introduced the `Anchored k-medoid`, package referred to as the `Ak-medoids`, which implements a medoid-based expectation maximisation (MEM) procedures within a classical k-mean clustering routine. The package includes functions that allow certain pre-processing of longitudinal data sets, prior to the clustering procedures. The potential application areas of `Ak-medoids` include the criminology, transport, epidemiology and brain imaging.

Source Code: Information:

Design and implementation

In longitudinal data clustering (LDA), studies have taken the advantages of the various functional characteristics of data in order to extract clusters of interest from longitudinal data sets. Examples include using the Fourier basis [@Tarpey2003] or the coefficients of the B-spline derivative estimates [@Boor1978; @Schumaker2007] of the datasets in order to anchor the clustering routines. Here, we develop an `Anchored k-medoids` (`Akmdeoids`) clustering package which employs the ordinary least square (OLS) trend lines of subjects in order to capture their long-term linear growths which may corresponds to theoretically meaningful developmental processes of the phenomenon over time. Particularly in criminology, such slowly changing trends may help to unravel certain place-based characteristics that drive crime-related events, such as street and gang violence, across a geographical space [@Griffith2004]. In related research, studies have deployed existing techniques, such as the k-means [@Curman2015; @Andresen2017] and GBTM [@Weisburd2004; @Chavez2009; @Bannister2017] techniques, which are more suited for spherical clusters [@GenoFali2010]. Moreover, the sensitivity of these techniques to short-term fluctuations and outliers in longitudinal datasets makes it more difficult for extracting clusters based on the underlying long-term growth over time.

The main clustering function in the `Akmdeoids` package implements a medoid-based expectation maximisation (MEM) procedures by integrating certain key modifications into the classical k-means routines. First, it approximates longitudinal trajectories using the ordinary least square regression (OLS) and second, anchors the initialisation process with medoid observations. It then deploys the medoids observations as new anchors for each iteration of the expectation-maximisation procedure [@Celeux1992], until convergence. In similar fashion as the classical k-means, the routine relies on distance-based similarity between vectors of observations and is scale invariant. This implementation ensures that the impacts of short-term fluctuations and outliers in the longitudinal dataset are minimised. The final groupings are projected on the raw trajectories to provide a clearer delineation of the long-term linear trends of trajectories. Given an `l` number of iterations, the computational complexity of the clustering routine is the same as that of a classical k-means algorithm, i.e. `O(lkn)`, where `k` is the specified number of clusters and `n`, the number of individual trajectories. The optimal number of clusters for a given data may be determined using the Calinski and Harabatz criterion [@Calinski1974] or the average silhouette [@Rousseeuw1987]. A full demonstration is provided in the package vignettes of how to deploy `Akmedoids` to examine long-term relative exposure to crime in micro-places. We encourage the use of the package outside of criminology, should it be appropriate.

Clustering and cluster representations

The main clustering function of akmedoids is the `akmedoids.clust`. The function generates directionally homogeneous clusters of a longitudinal data sets. For crime inequality studies, the package includes the `props` function for converting the absolute (or rate) measures of individual trajectories into relative measure over time. The `statPrint` function draws from the `ggplot2` library [@ggplot2] in order to visualise resulting clusters in either a line or a areal-stacked graph format, alongside basic cluster statistics.

Acknowledgment

We gratefully acknowledge the Economic and Social Research Council (ESRC), who funded the Understanding Inequalities project (Grant Reference ES/P009301/1) through which this research was conducted.

References