

Clustering Life Courses: A New Divisive Approach

Il Raggruppamento dei Corsi di Vita: un Nuovo Approccio Divisivo

Raffaella Piccarreta - Francesco C. Billari

IMQ, Università Bocconi, viale Isonzo 25, 20135 Milano, Italy

Riassunto: Obiettivo del lavoro è lo studio di corsi di vita individuali (carriera lavorativa e famiglia) rappresentati attraverso sequenze. Il problema del raggruppamento viene affrontato in un'ottica divisiva, utilizzando un algoritmo la cui logica attinge a quella della procedura CART, opportunamente estesa.

Keywords: sequence analysis, hierarchical clustering methods, life course analysis, work and family, CART.

1. Introduction

We apply a methodology for clustering individual life courses to data from the British Household Panel Survey (BHPS) on women born from 1960 to 1968. The basic idea is to develop clusters of women who follow similar life course paths during early adulthood. We follow 578 women over the age span 13-29 (cases with missing information are excluded). For each woman we build, on a monthly time scale (204 time points), a sequence-type representation of three life course domains: work (**w**), coresident partnership (**u**), both represented in a dichotomous manner, and fertility (4 possible fertility states ranging from 0 to 3 children and above). The work and partnership status can change in either direction at any time, while fertility are treated as irreversible events. Thus the number of children cannot decrease. In total this produces 16 different life course states: **0, u, w, wu, 1, 1u, 1w, 1wu, 2, 2u, 2w, 2wu, 3, 3u, 3w, 3wu** (a 3 indicates 3 children or more). Frequencies of single sequences are very low, so that it is necessary to obtain clusters to describe them. In Aassve et al. (2003), the problem of clustering sequences was solved by using hierarchical agglomerative methods, and some new criteria to meaningfully represent the obtained clusters were illustrated. In this paper we introduce a new divisive clustering strategy; a monothetic divisive algorithm was applied by Billari and Piccarreta (2001), but only to life course sequences fulfilling special properties (i.e. corresponding to non-repeatable events).

2. Clustering sequences: a divisive algorithm

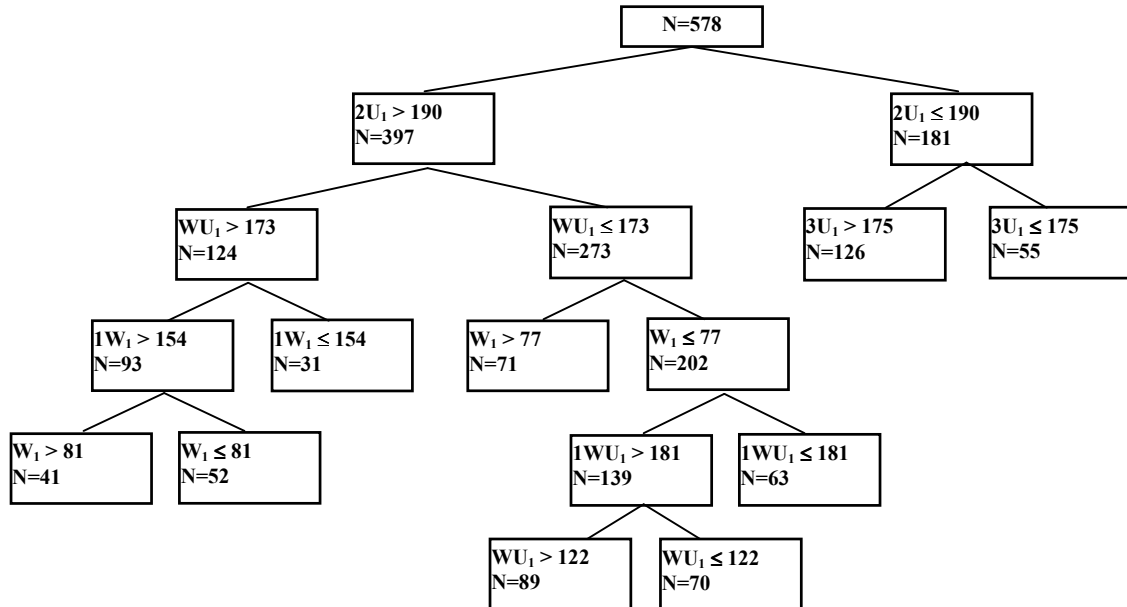
To apply standard (hierarchical) algorithms to life course data in a sequence-type representation it is necessary to define a distance (dissimilarity) matrix. To compute dissimilarities we use *Optimal Matching Analysis* (OMA), a widely accepted criterion introduced by Abbott (1995). The basic idea of OMA is to measure the dissimilarity between two sequences by properly quantifying the effort needed to transform one sequence into another. In the most elementary approach, three basic operations to

transform sequences are used: *insertion* (ι) - one state is inserted into the sequence - *deletion* (δ) - one state is deleted from the sequence - *substitution* (σ) - one state is replaced by another one. To each elementary operation a cost can be assigned. The distance between two sequences is defined as the minimum (total) cost of transforming one sequence into the other one. Thus the OMA-distance (which can be computed using TDA - Rohwer and Pötter, 2002) takes into account the entire sequences. To assign costs to transform one sequence into the other one we adopt a data-driven approach: substitution costs are inversely proportional to transition frequencies. In Aassve et al. (2003), on the basis of the OMA-distance matrix, the solution provided by Ward's algorithm was chosen. To better motivate our divisive-proposal, let us first analyze more in detail the characteristics of Ward's approach. Consider N sequences: a measure of the dispersion characterizing the whole sample is $\mathbf{T} = \sum_{i,j} d^2(i,j)$, $d^2(i,j)$ being the OMA-distance between the i -th and the j -th sequence. If the sample is divided into G clusters, the dispersion within the g -th group is $W_g = \sum_{(i,j) \in g} d^2(i,j)$; hence a measure of the within groups dispersion is $\mathbf{W}_G = \sum_g W_g$. The quantity $\mathbf{B}_G = \mathbf{T} - \mathbf{W}_G$ can be regarded as a synthesis of the distances among the G groups. A commonly used criterion to evaluate the adequacy of clustering is $M_G = R_G^2 = \mathbf{B}_G / \mathbf{T}$, the proportion of the total dispersion accounted for by the G clusters. If the whole sample is divided into G clusters, in an agglomerative-procedure ($G - 1$) clusters have to be obtained by joining two clusters into a single one. In Ward's algorithm the two clusters to be joined are selected by *minimizing* the quantity $\Delta M_{G-1} = (M_G - M_{G-1})$. If one refers to the same criterion in a divisive setting, starting from $G - 1$ clusters, G clusters have to be obtained by splitting a cluster. The cluster to be split is selected by *maximizing* $\Delta M_{G-1} = (M_G - M_{G-1})$. The main problem now is how to define the class of possible splits for a given group. In the present work, we consider the splits induced by the values of a set of (auxiliary) variables. In particular, a group g can be divided into two subgroups on the basis of the values of one of the variables, called the *splitting variable*. In particular, if A_q is a set of values assumed by X_q in g , a split s defined on the basis of A_q divides g into two subgroups $g_L = \{i: i \in g, x_{iq} \in A_q\}$ and $g_R = \{i: i \in g, x_{iq} \notin A_q\}$. Split s can be evaluated by considering $\Delta(g, s) = W_g - W_{g_L} - W_{g_R}$. The best split, $s^*(g)$, is selected in such a way that ΔM_{G-1} is maximum, $s^*(g) = \operatorname{argmax}_s \Delta(g, s)$. Hence both Ward's and this divisive algorithm aim at maximizing at each step M_G , even if in two opposite "directions". Ward's algorithm is more flexible, since the selection of the groups to be joined is not related to considerations about the levels of a particular variable. The divisive algorithm has the nice, and important, characteristic to lead to groups that are perfectly characterized and easy to be interpreted, at least in terms of the splitting variables.

In the problem of clustering sequences, the dissimilarity matrix is the OMA-distance matrix. A key problem is how to define the set of auxiliary variables to be utilized in selecting the splits. For this purpose, we consider a set of variables indicating the time of the first (second, third, and so on) passage in a particular state. For example $\mathbf{WU}_1$ is the variable indicating the month when an individual experienced state $\mathbf{WU}$ for the first time; more generally, while $\mathbf{WU}_k$ indicates the time of the k -th passage in state $\mathbf{WU}$. Before proceeding, we should point out that these auxiliary variables can not be directly used in a clustering procedure, since the (euclidean) distance defined on the basis of these variables is based on less information than that contained in sequences. Another

important consideration is the following: readers who are familiar with the CART procedure to build tree-classifiers to predict a response variable on the basis of a set of Q explanatory variables, will notice that there is a great similarity between this divisive algorithm and CART. The main difference is that in the present context the impurity within a node is not measured on the basis of the values of the response variable but, rather, on the basis of a distance/dissimilarity matrix defined on the basis of the response. This extension of CART to this situation has been recently introduced and studied by Piccarreta (2003). Following the CART reasoning, we obtained and pruned the *maximal* tree (see Breiman et al., 1984). Notice that, by referring to CART, one could select the proper number of groups by cross-validating the “tree” (this is an extension of the present work which is already under study and which is not extensively described due to space limitation). In the present context the number of clusters was chosen by referring to standard criteria (analysis of R^2 , ΔR^2). The 9-clusters partition was chosen, also to compare results with those obtained using Ward’s algorithm, described in Aassve et al. (2003).

Figure 1: *Tree representation of the 9-clusters solution*



3. A first interpretation and discussion

Given space constraints, we only add a basic interpretation of the results we obtained. The most important implication is that the rules illustrated in Figure 1 allow for an easy-to-read clustering and an immediate interpretation of groups. This is the most important, and yet unsolved, problem in the analysis of life courses as sequences.

We focus our interpretation on one peculiar group. The third group from the bottom left in Figure 1 is characterized by three splits in terms of the auxiliary variables: first, the passage from the state $2U$ happens no earlier than in month 190, say after about age 29 (note that these cases may also indicate right-censoring, that is we do not know whether women who follow this route will ever touch the state $2U$). In the second split, the passage from the state WU is experienced no earlier than in month 173, after about age

27. In the third split, the passage from the state 1w is experienced no later than in month 154 (before about age 26). The main feature of this group is then the *experience of single motherhood*; women who even experience state 1w have been mothers and workers without being in a union (either directly or as a consequence of union dissolution). Single motherhood is clearly a key socio-demographic and policy issue in the British context, and our algorithm automatically detects a group with such feature. This cluster contains 31 women (5.4% of the sample), and it is very heterogeneous as the average OMA distance within the cluster is 235.5, the highest value of all 9 groups, indicating that the experience of single motherhood characterizes women following before and after very different life paths. Nonetheless, the clarity of the tree produced by the divisive algorithm allows a clear interpretation for such group.

We also compared the groups obtained with this algorithm with the groups obtained using Ward's procedure in Aassve et al. (2003). The two procedures do not significantly overlap. We believe the divisive algorithm has a great advantage in terms of substantive interpretation, and this has implications for instance when one wants to obtain policy-related groupings.

References

- Aassve A., Billari F.C., Piccarreta R. (2003) Sequence Analysis of BHPS Life Course Data, *Book of Short Papers, CLADAG 2003, Classification and Data Analysis Group Italian Statistical Society*, Bologna: 22-24 Settembre 2003, 13-16.
- Abbott A. (1995) Sequence Analysis: New Methods for Old Ideas, *Annual Review of Sociology*, 21, 93-113.
- Billari F.C., Piccarreta R. (2001) Life courses as sequences: an experiment in classification via monothetic divisive algorithms, in *Advances in Classification and Data Analysis*, S. Borra, R. Rocchi, M. Vichi, M. Schader (Eds.), Springer Verlag, Berlin and New York, 351-358.
- Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984) *Classification and Regression Trees*, Wadsworth, Belmont, CA.
- Rohwer G., Pötter U. (2002), *TDA User's Manual*, Ruhr-Universität-Bochum.
- Piccarreta R. (2003) CART for Distance or Dissimilarity Matrices, *Studi Statistici* 80, Istituto di Metodi Quantitativi, Università Commerciale "L. Bocconi", Dicembre 2003.