ClusPath: a Temporal-driven Clustering to Infer Typical Evolution Paths

M.A. Rizoiu, J. Velcin, S. Bonnevay & S. Lallich
Australian National University & University Lumière Lyon

1 The problem

- Detect typical evolution paths of individuals through time.
- "Slow changing world" hypothesis: changes in the population are gradual and smooth, detect of long term trends.
- Allow the relations between phases to emerge from the data.

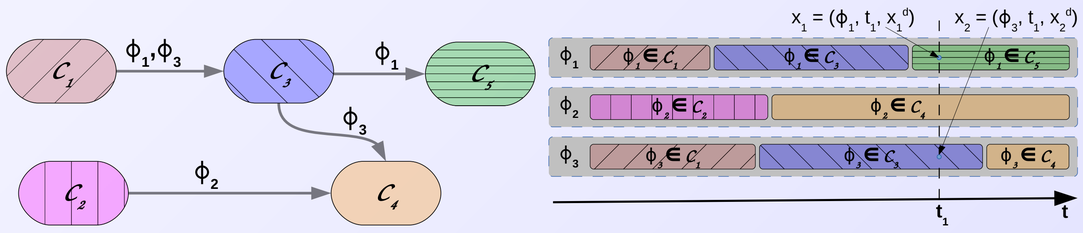

Figure 1: An example of a desired output, in which the evolutions of 3 entities (ϕ_1, ϕ_2 and ϕ_3) are described using 5 phases ($\mathcal{C}_i, i = 1, \ldots, 5$). For example, the evolution path of ϕ_3 is $\mathcal{C}_1 \rightarrow \mathcal{C}_3 \rightarrow \mathcal{C}_3$. **Left:** The graph structure of the evolution phases. The arcs between two phases $(\mathcal{C}_i, \mathcal{C}_j)$ are labeled with the entities presenting the transition $\mathcal{C}_i \rightarrow \mathcal{C}_j$. **Right:** The observations of the 3 entities are partitioned contiguously into the 5 phases.

2 A temporal clustering solution

- A temporal-aware constrained clustering algorithm, for which the resulted clusters serve as evolution phases.
- The relations between the evolution phases are inferred simultaneously with the partition.

3 objectives

Obj. 1 Construct clusters that are coherent in the temporal and in the descriptive space, using a temporal aware dissimilarity measure:

$$||x_i - x_j||_{TA} = 1 - \left(1 - \gamma_d \frac{||x_i^d - x_j^d||^2}{\Delta d_{max}^2}\right)\left(1 - \gamma_t \frac{||x_i^t - x_j^t||^2}{\Delta t_{max}^2}\right) .$$

Obj. 2 Segment, as contiguously as possible, the series of observations for each entity, using temporally-oriented soft must-link pair-wise constraints. Inflict the penalty $w(x_i, x_k)$ when the constraint between x_i and x_k is violated.

$$w(x_i, x_k) = \beta * e^{-\frac{1}{2}\left(\frac{||x_i^t - x_k^t||}{\delta}\right)^2}\left(1 - a_{j,l}^2\right)$$

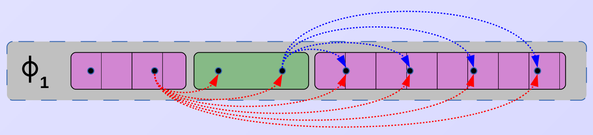

Figure 2: Example of forward-oriented must-link constraints violated by the given segmentation.

Obj. 3 Present smooth passages between phases on evolution paths, i.e., changes should come in small increments.

- The evolution phases are structured as an oriented graph: $a_{p,q}$ strength of link between $\mathcal{C}_p$ and $\mathcal{C}_q$.
- The strength of the link from $\mathcal{C}_p$ to $\mathcal{C}_q$ is proportional to the similarity of their prototypes μ_p and μ_q:

$$T_2 = \Sigma_{\mu_p \in \mathcal{M}} \Sigma_{\substack{\mu_q \in \mathcal{M} \\ p \neq q}} a_{p,q}^2 ||\mu_p - \mu_q||_{TA}$$

- The strength of the link from $\mathcal{C}_p$ to $\mathcal{C}_q$ is proportional to the number of entities that present a transition from $\mathcal{C}_p$ to $\mathcal{C}_q$:

$$inter_\phi(\mathcal{C}_p, \mathcal{C}_q) = 1 - \frac{|\{\phi_l \in \Phi | \mathcal{C}_p \xrightarrow{\phi_l} \mathcal{C}_q\}|}{|\Phi|}$$
$$T_3 = \Sigma_{\mu_p \in \mathcal{M}} \Sigma_{\substack{\mu_q \in \mathcal{M} \\ p \neq q}} a_{p,q}^2 inter_\phi^2(\mathcal{C}_p, \mathcal{C}_q)$$

The ClusPath algorithm

- K-Means inspired, iterates three update phases: a) recompute prototypes, b) assign observations to clusters and c) recompute adjacency matrix between phases.
- Evaluate resulted partition using four measures: **MDvar**, **Tvar** (descriptive and temporal variance), **ShaP** (contiguous segmentation) and **SPass** (smooth passage)

ShaP $= \frac{1}{|\Phi|} \Sigma_{\phi \in \Phi} \Sigma_{i=1}^{k} \left[-p_\phi(\mathcal{C}_i) \log_2(p_\phi(\mathcal{C}_i))\right]\left(1 + \frac{n_{ch} - n_{min}}{N-1}\right), \quad p_\phi(\mathcal{C}_i) = \Sigma_{\substack{x_j \in \mathcal{C}_i \\ x_j^\phi = \phi}} \frac{1}{N}$

SPass $= \Sigma_{\phi \in \Phi} \Sigma_{\substack{\mathcal{C}_i \xrightarrow{\phi} \mathcal{C}_j}}^{i,j \in 1,\ldots,k} \frac{||\mu_i - \mu_j||_{TA}}{n_{ch}}$

3 Two datasets

Comparative Political Dataset I
- 23 countries
- 60 years
- 207 political, social and economic variables

European Companies
- 836 companies
- 5 years
- 7 economic variables

Automatically setting parameters using an evolutionary technique

The six parameters of ClusPath (α, β, δ, $\lambda_1, \lambda_2, \lambda_3$) can be automatically tunned using an evolutionary technique, that approximates the Pareto front in the four-dimensional space of the measures. We chose the parameters that produce a balanced solution (closest to the ideal point $(0, 0, 0, 0)$).

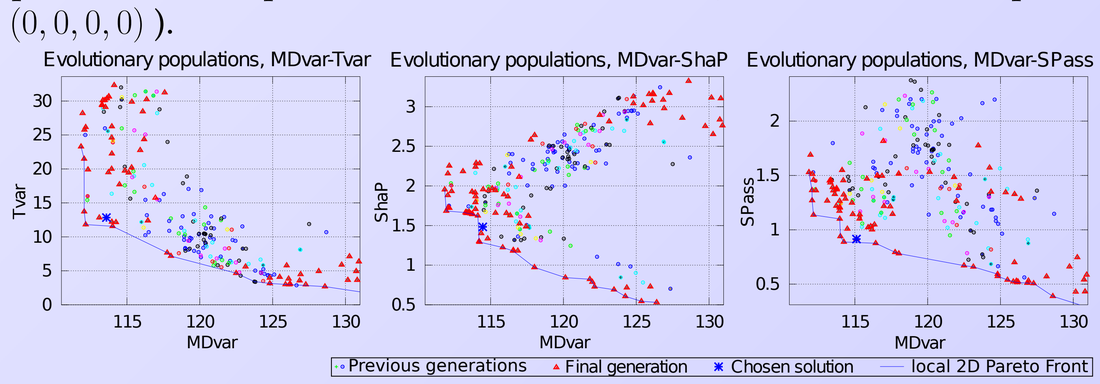

Figure 3: Typical example of execution of the evolutionary algorithm on `CPDS1`. The obtained 4-dimensional Pareto front is projected onto the (*MDvar*, *Tvar*) space (left), (*MDvar*, *ShaP*) space (middle) and (*MDvar*, *SPass*) space (right).

4 Detecting political systems

Typical evolution paths are taken by most entities (e.g. $\mathcal{C}_7 \longrightarrow \mathcal{C}_{11} \longrightarrow \mathcal{C}_{15}$). Their meaning emerges by studying the entities following them: $\mathcal{C}_5 \longrightarrow \mathcal{C}_{11}$ (Spain, Portugal and Greece) coincides with non-democratic regimes and $\mathcal{C}_4 (\longrightarrow \mathcal{C}_7) \longrightarrow \mathcal{C}_{11}$ (Denmark, Finland, Iceland, Norway and Sweden) maps onto the "Swedish Social and Economical Model".

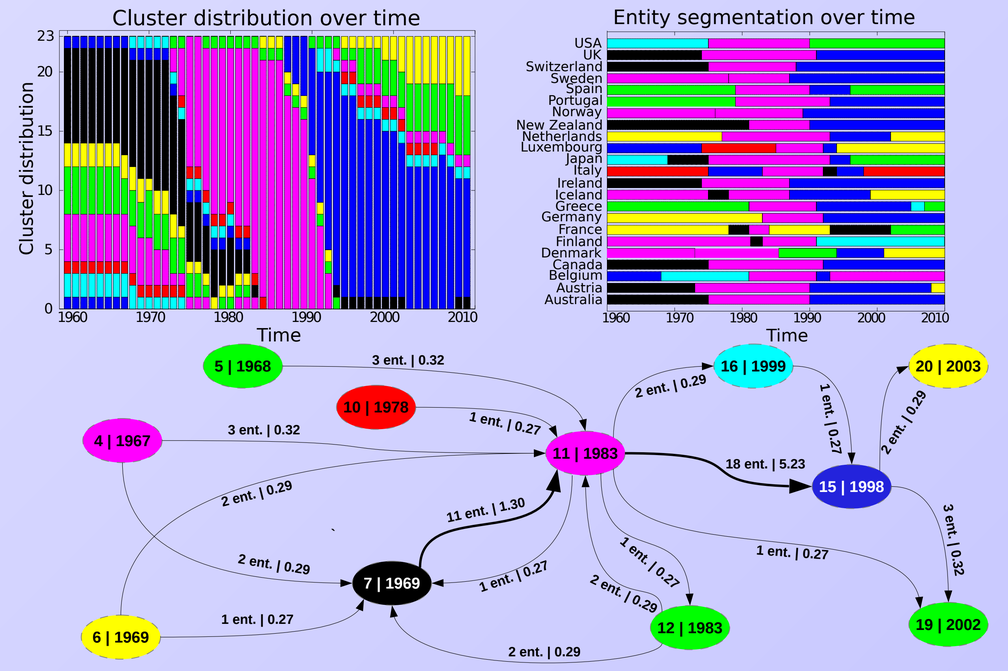

Figure 4: Typical evolution phases constructed by ClusPath on `CPDS1`, with 20 clusters. Number of entities in each phase per year (a), segmentation of entities over phases (b) and the phase evolution graph (c)

5 Detecting fiscal optimization

"Tax optimization" undertook by most companies in the `EC` dataset: most companies arrive to decrease significantly their tax rate, while increasing the net income.

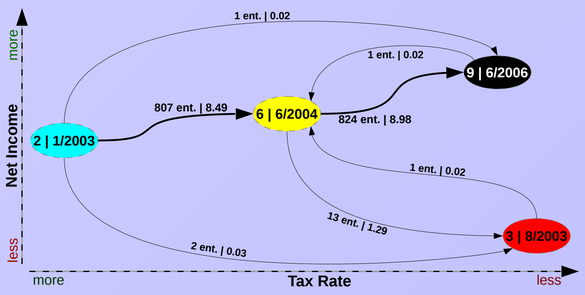

Figure 5: Typical evolution phases constructed by ClusPath on `EC`, with 10 clusters. The evolution graph is projected in the space *NetIncome*/*TaxRate*.