A Validation Methodology in Hierarchical Clustering

Fernanda Sousa Jorge Tendeiro

FEUP, Universidade do Porto

ASMDA 2005

Brest, 17-20 May

Introduction

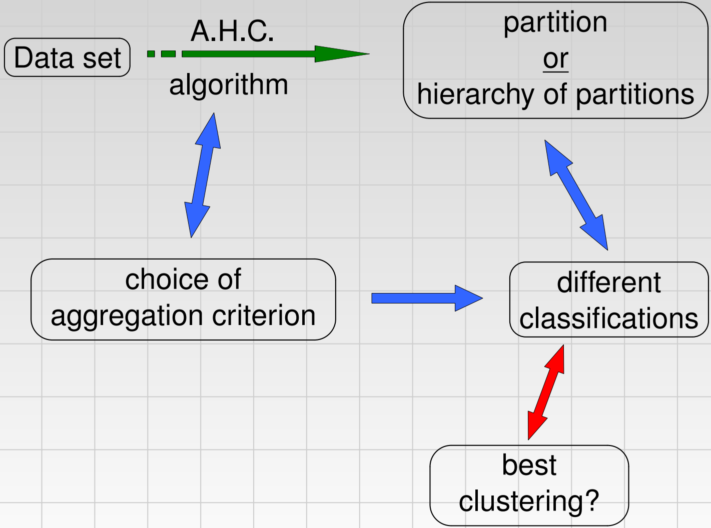
Area of research

Ascending Hierarchical Clustering (A.H.C.)

Presentation scheme

- validation in A.H.C.
  - comparison of clustering structures
  - random generation of dendrograms or ultrametric matrices
- methodology of validation
- an application
- conclusion and perspectives

Motivation


Data set
A.H.C.
algorithm
partition
or
hierarchy of partitions
choice of
aggregation criterion
different
classifications
best
clustering?

Some questions

Questions

- Is there a structure of the initial data? Is there a close relation between the initial and final structures?
- Which choice of comparison functions is to result into the best clustering?
- How can we assure that the division into several clusters suggested by the algorithm does not distort the structure of the initial data?
- Do the relations between the elements to classify lead to artificial clusters without real meaning?

Several contributions

- Bock. . .
- Gordon e Milligan. . .
- Lapointe e Legendre. . .
- Barthélemy *et* al. . .
- Bel Mufti. . .
- Hubert. . .

Validation Methodology in A.H.C.

Results of an A.H.C. method depend on. . .

- inicial data
- method used

Moreover. . .

The behaviour of an A.H.C. method is influenced by the structure of the data.

Main goal

Describe the performance of several A.H.C. methods, when applied to different types of data.

Useful tools

- comparison of clustering structures
- random generation of dendrograms

Comparison of clustering structures

Ordinal approach

Uses the ordenations of indexed values.

Idea

clustering structures (proximity matrix, hierarchy, partition)

↕

preordenations

“Pratical” consequence

Comparison of clustering structures transformed into comparison of preordenations.

Random generation of dendrograms

We want to randomly generate. . .

- topologies
- labels
- aggregation levels

Methods used

- uniform *sensu* Furnas (Furnas 1984)
  - Uniform (Sousa & Nicolau 2000)
  - Double Permutation (Lapointe & Legendre 1991)
  - RA (Podani 2000)
- not uniform
  - Shape Parameter (Sousa 2000)

Methodology

For a fixed number of elements to classify, consider the following steps:

Algorithm (1 of 2)

1. Generate a random dendrogram; the associated ultrametric matrix, M_0, will be taken as the (initial) dissimilarity matrix.
2. For each A.H.C. criterion to study: obtain a hierarchy H_0, and compare M_0 with H_0 (comparison $\mathcal{C}^1$).
3. Disturb matrix M_0 by settling a disturbance coefficient; this creates the dissimilarity matrix M_i. Compare M_0 with M_i (comparison $\mathcal{C}^2$).

$\vdots$

Methodology

Algorithm (2 of 2)

⋮

4. For each A.H.C. criterion to study: obtain a hierarchy H_i, compare M_i with H_i (comparison $\mathcal{C}^3$) and compare H_0 with H_i (comparison $\mathcal{C}^4$).
5. Repeat the steps 3. and 4. a great number of times for the same disturbance coefficient.
6. Repeat the steps 3. to 5. for different values of the disturbance coefficient.

Methodology

$$\begin{tikzcd} M_0 \arrow[r, leftrightarrow, "\mathcal{C}^1"] \arrow[d, leftrightarrow, "\mathcal{C}^2"'] & H_0 \arrow[d, leftrightarrow, "\mathcal{C}^4"] \\ M_i \arrow[r, leftrightarrow, "\mathcal{C}^3"'] & H_i \end{tikzcd}$$

Structures

- M_0: generated ultrametric matrix
- H_0: output of an A.H.C. applied to M_0
- M_i: ultrametric matrix after disturbance
- H_i: output of an A.H.C. applied to M_i

Methodology

$$\begin{tikzcd}
\boxed{M_0} \arrow[r, leftrightarrow, "\mathcal{C}^1"] \arrow[d, leftrightarrow, "\mathcal{C}^2"'] & \boxed{H_0} \arrow[d, leftrightarrow, "\mathcal{C}^4"] \\
\boxed{M_i} \arrow[r, leftrightarrow, "\mathcal{C}^3"'] & \boxed{H_i}
\end{tikzcd}$$

Comparisons

- $\mathcal{C}^1$: analyse a criterion behaviour when applied to ultrametric data
- $\mathcal{C}^2$: control the impact of the disturbance over the associated preordenations
- $\mathcal{C}^3$: analyse the ability of a criterion to recover a structure after disturbance
- $\mathcal{C}^4$: controls the robustness of the method

Options taken

Options

- number of elements to classify: 10
- three types of structures generated:
  - predominantly chain type trees (shape parameter method, p close to 0)
  - predominantly balanced trees (shape parameter method, p close to 0.5)
  - completely random trees (uniform method)
- several methods of A.H.C.:
  - classical approach (SL, CL, HMEAN, HMED)
  - VL approach (AVB, AVM, HVMED)
- different values of the disturbance coefficient
- coefficient of comparison used: Goodman-Kruskal
- comparisons $\mathcal{C}^1$, $\mathcal{C}^2$, $\mathcal{C}^3$, $\mathcal{C}^4$ analysed

Comparison $\mathcal{C}^1$: $M_0 - H_0$

T_{GK} values, uniform generation

	AVB	AVM	HVMED
mean	.725	.839	.885
median	.749	.913	.940
dispersion	.445	.094	.094

T_{GK} values, shape parameter, $p = .025$

	AVB	AVM	HVMED
mean	.632	.966	.970
median	.644	1	1
dispersion	.020	<0.001	<0.001

T_{GK} values, shape parameter, $p = 0.5$

	AVB	AVM	HVMED
mean	.851	.792	.844
median	.893	.851	.932
dispersion	.193	.018	.433

Comparison $\mathcal{C}^2$: $M_0 - M_i$

Utility

- the analyse of the T_{GK} values was useful to determine several disturbance values for $\mathcal{C}^3$ e $\mathcal{C}^4$ comparisons
- 4 values of disturbance δ were considered

Comparison $\mathcal{C}^3$: $M_i - H_i$

Median values of T_{GK}, uniform generation

δ	SL	CL	HMEAN	HMED	AVB	AVM	HVMED
.05	.965	.965	.968	.968	.610	.795	.841
.15	.681	.727	.748	.743	.598	.581	.652
.25	.581	.618	.658	.657	.561	.496	.569
.5	.404	.456	.527	.523	.421	.341	.418

Comparison $\mathcal{C}^3$: $M_i - H_i$

Median values of T_{GK}, shape parameter, $p = .025$

δ	SL	CL	HMEAN	HMED	AVB	AVM	HVMED
.05	.938	.966	.959	.959	.543	.935	.933
.15	.766	.773	.808	.805	.530	.748	.747
.25	.656	.641	.717	.718	.516	.665	.662
.5	.438	.479	.556	.548	.408	.431	.449

Median values of T_{GK}, shape parameter, $p = .5$

δ	SL	CL	HMEAN	HMED	AVB	AVM	HVMED
.05	.948	.952	.953	.953	.844	.586	.805
.15	.721	.728	.764	.754	.688	.566	.673
.25	.640	.620	.690	.687	.611	.525	.597
.5	.420	.470	.540	.536	.432	.347	.431

Comparison $\mathcal{C}^4$: $H_0 - H_i$

Median values of T_{GK}, uniform generation

δ	SL	CL	HMEAN	HMED	AVB	AVM	HVMED
.05	.990	.993	.992	.991	.944	.958	.970
.15	.954	.927	.949	.920	.881	.904	.869
.25	.778	.709	.819	.789	.748	.664	.679
.5	.472	.353	.522	.493	.442	.271	.345

Comparison $\mathcal{C}^4$: $H_0 - H_i$

Median values of T_{GK}, shape parameter, $p = .025$

δ	SL	CL	HMEAN	HMED	AVB	AVM	HVMED
.05	.978	.971	.979	.974	.828	.959	.941
.15	.930	.810	.914	.887	.759	.909	.868
.25	.878	.611	.810	.780	.560	.859	.820
.5	.617	.251	.523	.479	.341	.596	.560

Median values of T_{GK}, shape parameter, $p = .5$

δ	SL	CL	HMEAN	HMED	AVB	AVM	HVMED
.05	.989	.986	.990	.980	.984	.977	.938
.15	.890	.859	.922	.877	.898	.762	.775
.25	.833	.701	.814	.790	.800	.630	.641
.5	.502	.399	.566	.523	.559	.347	.446

Some conclusions

From the application we can say that. . .

- VL methods have more difficulty to recover the inicial structure data than classical methods
- classical and VL methods are equally robust (similar ability to resist to disturbances of the data)
- behaviour of VL methods:
  - *AVB*: better with balanced trees
  - *AVM*: better with chain trees
  - *HVMED*: it's the one which resists most to the variation of data structure
- behaviour of classical methods:
  - *SL*: works well with chain trees; very robust
  - *CL*: works well with balanced trees
  - *HMEAN* and *HMED*: similar behaviour in all situations analysed

General conclusions and perspectives

Topics

- need to validate clustering results
- the behaviour of a clustering method strongly depends on the kind and intensity of the data structure
- simulation studies are very useful in this area, since theoretical studies are extremely difficult
- lead studies with different options

References

Frank, O.; Svensson, K. (1981)
On Probability Distributions of Single-Linkage Dendrograms
Journal of Statistical Computation and Simulation, 12:121–131

Lapointe, F. J.; Legendre, P. (1991)
The Generation of Random Ultrametric Matrices Representing Dendrograms
Journal of Classification, 8:177–200

Sousa, F. (2000)
Novas Metodologias e Validação em Classificação Hierárquica Ascendente
Dissertação de Doutoramento, Universidade Nova de Lisboa