Scan

Schaber

paper

add to 2 sets

AS121
A654

Metrika, Volume 27, 1980, page 127–132. ©*Physica-Verlag, Vienna.*

Hierarchical Analysis: Classification with Ordinal Object Dissimilarities

By *M. Schader*, Karlsruhe[1])

Summary: Lerman [1970] has demonstrated, that the dissimilarity indices normally used in data analysis are identical up to strictly monotone transformations $f: \mathbf{R}_+ \to \mathbf{R}_+$ if the data are nominal and each set of attribute scores is finite.

In that case he proposes to use a preorder between pairs of objects to express similarity or dissimilarity, in order to avoid inconsistent classification results that might occur, if clustering schemes which are not monotone invariant are applied to a quantitative index. Here it is shown, how a hierarchy on the objects can be calculated, if such a preorder relation is given.

1. Dissimilarity and Ultradissimilarity

Let X be a finite set and $\lesssim$ a dissimilarity (-relation) on X, that is

$\lesssim$ is a total preorder relation
in $Y := \{\{x, y\} \mid x, y \in X \text{ and } x \neq y\}$.

$\{w, x\} < \{y, z\}$ signifies, that the objects y and z are more different than w and x. For $\lesssim$ to be nontrivial, we assume that $|X| \geqslant 3$.

It is desired to construct a hierarchy on X, that is a family $P_0, P_1, \ldots, P_q$ of partitions of X with

$P_0 = \{\{x\} \mid x \in X\}, P_q = \{X\}$ and for $i \in \{0, 1, \ldots, q-1\}$
P_i is strictly finer than P_{i+1} (i.e. $P_i \neq P_{i+1}$
and $\forall A \in P_i \;\; \exists B \in P_{i+1}$ such that $A \subset B$).

This hierarchy has to represent the given object dissimilarities "as accurately as possible".

Defining an ultradissimilarity on X as a dissimilarity $\lesssim'$ on X that satisfies

$$\{x, z\} \lesssim' \{x, y\} \quad \text{or} \quad \{x, z\} \lesssim' \{y, z\}$$

for all $x, y, z \in X$ that are pairwise distinct, we can establish a function φ from the set of ultradissimilarities on X into the set of hierarchies on X:

---

[1]) *M. Schader*, Institut für Entscheidungstheorie und Unternehmensforschung, Universität Karlsruhe, D–7500 Karlsruhe.

If $\lesssim'$ is a given ultradissimilarity on X, let $C_1, C_2, \ldots, C_q$ be the equivalence classes with respect to $\sim'$ and assume that these classes are numbered such that $i<j$, $\{w, x\} \in C_i$ and $\{y, z\} \in C_j$ yields $\{w, x\} <' \{y, z\}$. Then the relations S_i

$$x S_i y :\Leftrightarrow x = y \text{ or } \{x, y\} \in C_1 \cup C_2 \cup \ldots \cup C_i \quad i \in \{1, \ldots, q\}$$

are equivalence relations in X, and for the corresponding partitions $P_1, \ldots, P_q$ of X P_i is strictly finer than P_{i+1}. Adjoining $P_0 = \{\{x\} \mid x \in X\}$ we obtain the hierarchy $P_0, \ldots, P_q =: \varphi(\lesssim')$.

φ is clearly an injection and furthermore, for any given hierarchy $P_0, \ldots, P_q$ on X there exists an ultradissimilarity $\lesssim'$ on X with $\varphi(\lesssim') = P_0, \ldots, P_q$ namely the relation defined by

$$\{w, x\} \lesssim' \{y, z\} :\Leftrightarrow \min\{i \mid \exists A \in P_i \text{ and } w, x \in A\} \leqslant \min\{i \mid \exists A \in P_i \text{ and } y, z \in A\}$$

(Hence $C_1 \cup \ldots \cup C_i = \{\{x, y\} \mid x, y \in X, x \neq y \text{ and } \exists A \in P_i \text{ with } x, y \in A\}$, thus $x S_i y \Leftrightarrow x = y$ or $\exists A \in P_i$ with $x, y \in A$ and therefore $\varphi(\lesssim') = P_0, \ldots, P_q$). In other words, φ is a bijection.

This allows us to restrict further considerations to the set $D(X)$ of dissimilarities on X:
Instead of calculating a hierarchy $P_0, \ldots, P_q$ on X, we construct the corresponding element $\lesssim'$ of the set $U(X)$ of ultradissimilarities on X (which is a subset of $D(X)$).

2. The Semilattices $D(X)$ and $U(X)$

$D(X)$, the (finite) set of dissimilarities on X can be ordered by an order relation R, setting for $\lesssim_1, \lesssim_2 \in D(X)$

$$\lesssim_1 R \lesssim_2 :\Leftrightarrow G_{\lesssim_1} \subset G_{\lesssim_2}.$$

Here $G_{\lesssim}$ denotes the graph of $\lesssim$, i.e. the set $\{(a, b) \mid a, b \in Y \text{ and } a \lesssim b\}$, and thus R is the order normally used for the comparison of relations.

$D(X)$, endowed with R, is a join semilattice, because any two elements $\lesssim_1$, $\lesssim_2 \in D(X)$ have a supremum in $D(X)$. The graph of $\sup\{\lesssim_1, \lesssim_2\}$ is the transitive closure of $G_{\lesssim_1} \cup G_{\lesssim_2}$, that is the smallest (with resp. to $\subset$) transitive graph $G \subset Y^2$ which contains $G_{\lesssim_1} \cup G_{\lesssim_2}$.

$U(X)$ is a sub-semilattice of $D(X)$, since $\lesssim_1, \lesssim_2 \in U(X)$ implies $\sup\{\lesssim_1, \lesssim_2\} \in \in U(X)$.

In addition, $D(X)$ is upper semimodular and therefore graded [cf. *Barbut/Monjardet; Birkhoff*], for example by

$$g(\lesssim) := |X|(|X|-1)/2 - |Y_{/\sim}|,$$

where $Y_{/\sim}$ designates the quotient set of Y by $\sim$, so that $|Y_{/\sim}|$ is the number of equivalence classes with respect to $\sim$.

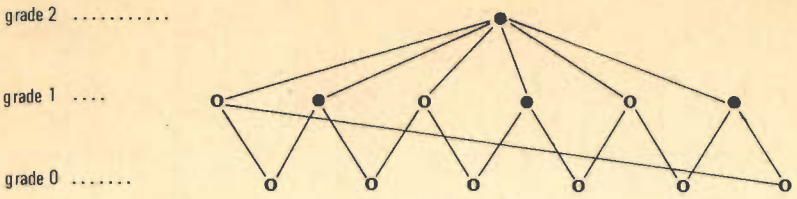

Fig. 1: The semilattice $D(X)$ if $|X| = 3$. Elements of $D(X) - U(X)$ resp. $U(X)$ are characterized by ○ resp. ●.

Using the Stirling numbers of the second kind $S(i, j)$, it is possible to compute the number of dissimilarities resp. ultradissimilarities on X that have grade 0, 1, 2 etc.

If $n := |X|$ and $m := n(n-1)/2 = |Y|$ there are [cf. *Barbut/Monjardet*]

$$(m-k)!\, S(m, m-k) \qquad 0 \leqslant k \leqslant m-1$$

dissimilarities of grade k.

Denoting by $N_k(l)$ (where $l \geqslant 3$ and $1 \leqslant k \leqslant l-1$) the number of ultradissimilarities with grade $l(l-1)/2 - k$ on a set that has cardinality l, and by $N_{kj}(l)$ the number of these ultradissimilarities having j classes ($k \leqslant j \leqslant l-1$) with respect to S_1 (S_1 defined as in 1.), we get

$$N_k(l) = \sum_{j=k}^{l-1} N_{kj}(l).$$

On the other hand, there are $S(l, j)$ ways to partition a set with cardinality l into j classes, so that

$$N_{kj}(l) = S(l, j)\, N_{k-1}(j).$$

It follows that, starting with $N_1(l) = 1$ for all l, we can find $N_k(l)$ and esp. $N_k(n)$ recursively according to

$$N_k(l) = \sum_{j=k}^{l-1} S(l, j)\, N_{k-1}(j)$$

[cf. hereto the computation of the total number of hierarchies on a finite set in *Lerman*].

$n = 3, 4, 5$ yields the following results:

	n = 3		n = 4		n = 5	
grade	number diss.	number ultra-diss.	number diss.	number ultra-diss.	number diss.	number ultra-diss.
0	6	-	720	-	3,628,800	-
1	6	3	1,800	-	16,329,600	-
2	1	1	1,560	-	30,240,000	-

grade	n = 3		n = 4		n = 5	
	number diss.	number ultra-diss.	number diss.	number ultra-diss.	number diss.	number ultra-diss.
3			540	18	29,635,200	-
4			62	13	16,435,440	-
5			1	1	5,103,000	-
6					818,520	180
7					55,980	205
8					1,022	50
9					1	1
Σ	13	4	4,683	32	102,247,563	436

A6541 and A5121

3. Approximation of Dissimilarities by Ultradissimilarities

If a hierarchy $P_0, \ldots, P_q$ on X has to be derived from a given dissimilarity $\lesssim$ on X, it is plausible to search for an element $\lesssim'$ of $U(X)$ that lies "close" to $\lesssim$ within the meaning of the structure described in section 2.

Since $D(X)$ is finite and a graded, upper semimodular join-semilattice, $d: D(X)^2 \to \mathbf{R}_+$, with

$$d(\lesssim_1, \lesssim_2) := 2g(\sup\{\lesssim_1, \lesssim_2\}) - g(\lesssim_1) - g(\lesssim_2)$$

is a distance function on $D(X)$ [cf. *Comyn/van Dorpe*] and respects the semilattice-order R in the way that

i) $\lesssim_1 R \lesssim_2 R \lesssim_3$ implies $d(\lesssim_1, \lesssim_3) = d(\lesssim_1, \lesssim_2) + d(\lesssim_2, \lesssim_3)$

ii) $d(\lesssim_1, \lesssim_2) = d(\lesssim_1, \sup\{\lesssim_1, \lesssim_2\}) + d(\lesssim_2, \sup\{\lesssim_1, \lesssim_2\})$

hold for any given $\lesssim_1, \lesssim_2, \lesssim_3 \in D(X)$.

With this distance function d (and $\lesssim \in (D(X) - U(X))$ given) we now search for $\lesssim' \in U(X)$ that minimizes $d(\lesssim, \lesssim')$.

Proposition

$\lesssim'$ is the least element (with respect to R)
of the set $M = \{\lesssim^* \mid \lesssim^* \in U(X) \text{ and } \lesssim R \lesssim^*\}$.

Proof: Note first, that $\lesssim'$ is an element of M. For if $\lesssim_1 \in D(X)$ and $\lesssim_2 \in U(X)$ then $\sup\{\lesssim_1, \lesssim_2\} \in U(X)$, and therefore $\lesssim' R \lesssim$ implies $\sup\{\lesssim, \lesssim'\} = \lesssim \in U(X)$. If on the other hand $\lesssim'$ and $\lesssim$ are incomparable, then $d(\lesssim, \sup\{\lesssim, \lesssim'\}) \leqslant d(\lesssim, \lesssim')$. Furthermore, $\lesssim_1, \lesssim_2 \in M$ together with $\lesssim_1 R \lesssim_2$ implies $d(\lesssim, \lesssim_1) \leqslant d(\lesssim, \lesssim_2)$. Hence the least element of M (when it exists) has minimal distance to $\lesssim$.

Now, if $\lesssim_1, \lesssim_2$ are elements of M, they have an infimum (with the graph $G_{\lesssim_1} \cap G_{\lesssim_2}$) which belongs to M, and since M is finite $\inf M$ exists and is an element of M. In other words, $\inf M$ is the least element of M. This completes the proof.

Thus, starting with $\lesssim$, we can calculate $\lesssim'$ by joining equivalence classes (with respect to $\sim$) if we find $x, y, z \in X$ that are pairwise distinct and do not satisfy the ultradissimilarity condition of section 1.:
If $\{x, y\} \lesssim \{y, z\} < \{x, z\}$ (resp. $\{y, z\} \lesssim \{x, y\} < \{x, z\}$) we set
$\{x, y\} \lesssim' \{y, z\} \sim' \{x, z\}$ (resp. $\{y, z\} \lesssim' \{x, y\} \sim' \{x, z\}$).

4. Algorithm and Example

To simplify the algorithmic description let $X = \{1, \ldots, n\}$ and let $\lesssim$ be defined by the equivalence classes with respect to $\sim$ viz. $C_1, C_2, \ldots, C_q$ which are again numbered as in section 1.

We calculate the classes with respect to $\sim'$:

Step 1

Set $x := 1,\ y := 2,\ z := 3$.

Step 2

Compute $i, j, k \in \{1, \ldots, q\}$ so that $\{x, y\} \in C_i$, $\{y, z\} \in C_j$ and $\{x, z\} \in C_k$.
Set $i_1 := \inf\{\sup\{i, j\}, \sup\{i, k\}, \sup\{j, k\}\}$ and $i_2 := \sup\{i, j, k\}$.
If $i_1 = i_2$ go to Step 3.
Set $C_{i_1} := \overset{i_2 - i_1}{\underset{\nu=0}{\cup}} C_{i_1+\nu}$.
If $i_2 < q$ set $C_{i_1+\nu} := C_{i_2+\nu}$ for $\nu \in \{1, 2, \ldots, q - i_2\}$.
Set $q := q - i_2 + i_1$.

Step 3

If $q = 1$ STOP: $\sim'$ has only one class C_1.
If $z < n$ set $z := z + 1$ and go to Step 2.
If $y < n - 1$ set $y := y + 1, z := y + 1$ and go to Step 2.
If $x < n - 2$ set $x := x + 1, y := x + 1, z := x + 2$ and go to Step 2.
STOP : $\sim'$ has the classes $C_1, \ldots, C_q$.

If for example $X = \{1, \ldots, 6\}$ and $\lesssim$ is the relation

$\{1,6\} < \{1,4\} \sim \{4,6\} < \{2,3\} < \{1,3\} \sim \{5,6\} < \{1,2\} \sim \{2,4\} \sim \{3,4\} \sim \{2,5\}$
$\sim \{3,5\} \sim \{4,5\} \sim \{2,6\} < \{3,6\} \sim \{1,5\}$

then the algorithm starts with $q = 6$ and the classes

$C_1 = \{\{1,6\}\}$
$C_2 = \{\{1,4\}, \{4,6\}\}$
$C_3 = \{\{2,3\}\}$
$C_4 = \{\{1,3\}, \{5,6\}\}$

$C_5 = \{\{1,2\}, \{2,4\}, \{3,4\}, \{2,5\}, \{3,5\}, \{4,5\}, \{2,6\}\}$
$C_6 = \{\{3,6\}, \{1,5\}\}$

and stops with $q = 4$ and

$C_1 = \{\{1,6\}\}$
$C_2 = \{\{1,4\}, \{4,6\}\}$
$C_3 = \{\{2,3\}\}$
$C_4 = \{\{1,2\}, \{1,3\}, \{1,5\}, \{2,4\}, \{2,5\}, \{2,6\}, \{3,4\}, \{3,5\}, \{3,6\}, \{4,5\}, \{5,6\}\}$

The corresponding hierarchy is shown in figure 2.

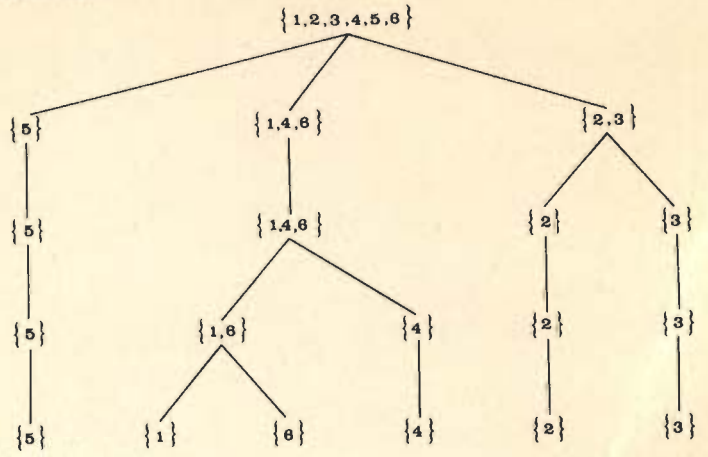

Fig. 2

References

Barbut, M., and *B. Monjardet*: Ordre et Classification 1, 2. Paris 1970.
Birkhoff, G.: Lattice Theory (3.ed.). American Mathematical Society, Providence 1973.
Comyn, G., and *J.C. van Dorpe*: Valuation et semi-modularité dans les demi-treillis. Math. Sci. hum. 56, 1976, 63–75.
Lerman, I.C.: Les Bases de la Classification Automatique. Paris 1970.

Received June 2nd, 1978
(revised version September 21, 1978)