Continuity!
To Be or Not to Be!
Is that even the question?

Melvin F. Janowitz

DIMACS, Rutgers University
Piscataway, NJ 08854

CS Meeting on June 16-18, 2011
at Carnegie Mellon University

Background material

P is a finite nonempty set of objects to be clustered
Dissimilarity coefficient (DC) $d : P \times P \rightarrow \Re_0^+$

$\bullet\, d(a,b) = d(b,a) \qquad \bullet\, d(a,a) = 0$

d is an ultrametric if also

$\bullet\, d(a,b) \leq \max\{d(a,c), d(b,c)\}$ for all $a, b, c \in E$.

Threshhold relation $Td(h)$ at level h for the DC d is the reflexive symmetric relation defined by $Td(h) = \{(a,b) : d(a,b) \leq h\}$

Fact:d is an ultrametric if and only if $Td(h)$ is an equivalence relation for all h.

d can be recaptured from Td since

$$d(a,b) = \bigwedge\{h : (a,b) \in Td(h)\}.$$

A Dilemma

Often times cluster methods are applied to data having only ordinal significance. Only care about whether $d(a, b) < d(x, y)$, and not about the actual values.

Now continuity *does* worry about how close $d(a, b)$ is to $d(x, y)$. So why even think about continuity?

Evolutionary biologists often assumed the existence of a true clustering associated with their input data (the evolutionary history of the given organisms), but that the data itself had in it some small errors caused by miscodings, improper characters, etc. Continuous cluster methods help put a bound on the resulting output errors.

Monotone Equivariance F is monotone equivariant (ME) if $F(\theta d) = \theta F(d)$ for all order automorphisms on $\Re_0^+$. When faced with ordinal data, it seems natural to at least use ME cluster methods, and avoid taking averages.

Order equivalence

Let $d \in D(P)$, the DCs on P. Let
image $(d) = [h_0 = 0 < h_1 < \cdots < h_t < h_{t+1}]$
with threshhold relations $R_0 \subset R_1 \subset \cdots R_t \subset R_{t+1} = P \times P$.
Then $\sigma(d) = \{R_0, R_1, \ldots R_t\}$ together with the image of d completely specifies d. Write $d \sim d'$ if $\sigma(d) = \sigma(d')$. Then $\sim$ is an equivalence relation on $D(P)$. Idea due to Robin Sibson (1972).

Say that F preserves order equivalence if $d \sim d'$ implies $F(d) \sim F(d')$. Say that cluster methods F, G are order similar if $F(d) \sim G(d)$ for all d.

Thus: $d \sim d' \implies d = \theta(d')$ for some order automorphism θ. Notion of $\sim$ ignores levels and only considers threshhold relations.

Defn: F compresses info if $F(d)$ cannot have more threshhold relations than does d.

Fact: If F preserves order equivalence then F is order similar to an ME cluster method iff F compresses informaton.

Single-Linkage Clustering

Recall: d is an ultrametric iff $Td(h)$ is an equivalence relation for all h.

Defn: A cluster method is often taken to be a transformation $d \mapsto Fd$ from a DC to an ultrametric.

Turns out we may define a cluster method F by

$$[T(Fd)](h) = \gamma([Td])(h) \text{ for all } h$$

where γ denotes the transitive closure. This is single-linkage clustering (SL).

With this definition of cluster method, SL is the unique method that is idempotent and isotone, and whose image is the set of all ultrametrics on P.

- idempotent: $F = F \circ F$.
- isotone: $d_1 \leq d_2 \implies F(d_1) \leq F(d_2)$.

$d_1 \leq d_2$ means $d_1(x,y) \leq d_2(x,y)$ for all $x, y \in P$.

How did Mel get into this?

Article: Reconstructing the history and geography of an evolutionary tree by D.Sankoff,
American Mathematical Monthly 79, 1972, pp. 596-603

Book: Mathematical Taxonomy by N. Jardine and R. Sibson,
Wiley, New York, 1971.

In this book, a number of desirable axioms are presented that cluster axioms might satisfy. SL is the unique cluster method that satisfies them all, and continuity is one of the axioms.

Huge controversy erupted in the literature, largely over continuity. Curious that continuity should after all be a consequence of some ordinal assumptions.

So: What is going on?

Flat cluster methods Cluster method F is flat if there is a mapping κ on relations such that $[T(Fd)](h) = \kappa(Td)(h)$ for all h.

Fact: Flat $\Longrightarrow$ isotone and ME. Converse fails!

Theorem: For ME cluster functions F, the following are equivalent:

- F is flat
- F is continuous
- F is right continuous
- $F(\theta d) = \theta F(d)$ for all 0-preserving isotone mappings θ on $\Re_0^+$.

This characterizes continuity in the presence of ME. So continuity is not all that closely tied to SL clustering. And continuity seems to be a consequence of assumptions that do not involve a metric on dissimilarities.

Problem: ME clustering assumes actual values of $d(a, b)$ versus $d(x, y)$ not important, only whether one is smaller than the other. But continuity worries about the actual values of the differences of values of pair of DCs. Is there some other property shared by continuous cluster methods?

Current project

Mesh Width $\mu(d)$

Image of d: $0 = h_0 < h_1 < \cdots < h_t < h_{t+1}$

$\mu(d) = \frac{1}{2} \min\{h_i - h_{i-1} : 1 \leq i \leq t+1\}$

Distance between DCs d and d'

$\Delta_0(d, d') = \max\{|d(a,b) - d'(a,b)| : a, b \in P\}$.

Theorem If $\Delta_0(d, d') < \mu(d)$ then

$$(1)\ d(a,b) < d(x,y) \text{ implies } d'(a,b) < d'(x,y)$$

Denote the assertion of equation (1) by the symbol $d \preceq d'$ (read d partially equivalent to d'). If both $d \preceq d'$ and $d' \preceq d$, we have the concept that has been denoted by $d \sim d'$: order equivalence. $\sim$ is an equivalence relation on $D(P)$.

Theorem $d \preceq d'$ is equivalent to every threshhold relation of d being a threshhold relation of d'.

Fundamental Fact: Let $d \preceq d'$. If $0 < \varepsilon < \mu(d)$, there exists d'' such that $d'' \sim d'$ and $\Delta_0(d, d'') < \varepsilon$. Thus if $[[d']]$ is the equivalence class of d' under $\sim$, then $d \preceq d'$ iff there exists a sequence d_n of $[[d']]$ with limit d.

More on Current project

Fact: If $d_n \to d$, there exists N such that $n \geq N$ implies $d \preceq d_n$.

Theorem Let F preserve order equivalence then F continuous forces F to preserve $\preceq$, but converse fails.

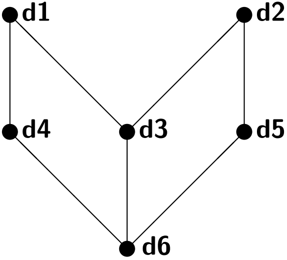
A natural setting more general than ME is to work with the equivalence classes of $\sim$. The notion of $\preceq$ induces a natural poset structure. Define $[[d_1]] \leq [[d_2]] \iff d_1 \preceq d_2$. Associate with $[[d]]$ the pair (k, W) where W is the weak order associated with d and k is 0 or 1 to denote whether $Td(0) = R_\emptyset$. This produces a structure called a semiBoolean algebra. See Janowitz, *On the semilattice of weak orders of a set*, Mathematical Social Sciences, 1985,**8**, 229-239. The poset of weak orders is characterized therein.

This leads one naturally to investigate cluster algorithms whose input is a weak order on the set of two element subsets of the finite set P of objects. $\preceq$ yields the dual of the usual ordering of weak orders. A key item is that $(0, W) < (1, W)$. Possible algorithms are described in Janowitz, *Monotone equivariant cluster methods*, SIAM Journal Applied Math, 1979, **37**, 148-165. The connection with continuity and order equivalence was not made there. The idea now is to investigate the theoretical background for DCs whose input and output are weak orders.

A weak order is a relation W that is
reflexive (xWx for all x);
symmetric(xWy implies yWx);
and complete (for any x, y, xWy or yWx).
With $P = \{a, b, c\}$ let $x = ab$, $y = ac$, $z = bc$ we are in the set of weak orders on $\{x, y, z\}$. There are 13 such weak orders.

Proper threshholds involving only x, y

	x	y	z	threshhold relations
$d1$	1	2	3	(x) (x, y) (x, y, z)
$d2$	2	1	3	(y) (x, y) (x, y, z)
$d3$	1	1	2	(x, y) (x, y, z)
$d4$	1	2	2	(x) (x, y, z)
$d5$	2	1	2	(y) (x, y, z)
$d6$	1	1	1	(x, y, z)



Things to do!

$\preceq$ was introduced in earlier talks, as well as in Chapter 4 of the text Ordinal and Relational Clustering by Janowitz.

- investigate further properties of $\preceq$.
- Investigate conditions in this model that cluster algorithms might satisfy.
- Develop computer implementations that can accommodate large data sets.
- Allow for data that has some numerical as well as just ordinal validity. A start was made toward this in a joint paper with Schweizer, *Ordinal and Percentile Clustering*, Math. Social Sciences, 1989, **18**, 135–186.