Continuous Ordinal Clustering: A Mystery Story[1]

Melvin F. Janowitz
DIMACS, Rutgers University
Piscattaway, NJ 07641

The Background

In their book "Mathematical Taxonomy", N. Jardine and R. Sibson presented a model for clustering algorithms that only allowed one feasible algorithm that produced an ultrametric output: single-linkage clustering. Among other things they assumed two axioms:

1. Clustering algorithms should be continuous.
2. Clustering algorithms should not be concerned with values of dissimilarities – only whether one value is larger or smaller than another.

But how can this be? The first condition involves the consideration of what happens when objects are close together. The second condition tells us to ignore closeness. This is a puzzle to be unravelled.

Definitions

The terminology in the area is not universal, so let's clarify the terms.

Input Data This is a finite nonempty set P of objects to classify. Each object has associated with it a set of numerical, binary, or nominal attributes.

Output Data A partition of P or an indexed nested sequence of partitions, the top one having a single class.

Intermediate step Convert the attribute data into a dissimilarity coefficient (DC). A DC on P is a mapping $d : P \times P \mapsto \Re_0^+$ such that

(1) $d(a, b) = d(b, a) \geq 0$

(2) $d(a, a) = 0$ for all $a \in P$.

d is *definite* if also

(3) $d(a, b) = 0$ implies $a = b$ in the sense that they are identical.

[1]Note: The present work has different goals and was done independently of the paper by O. Gascuel and A. McKenzie, *Performance Analysis of Hierarchical Clustering*, Journal of Classification, **11**, 2004, pp. 3-18, though there is some overlap of ideas.

d is an *ultrametric* if it satisfies (1), (2) and the ultrametric inequality

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

The DCs are ordered by the rule $d_1 \leq d_2 \iff d_1(a,b) \leq d_2(a,b)$ for all $a, b \in P$. The smallest DC is then given by $\underline{0}$ which is defined by $\underline{0}(a,b) = 0$ for all $a, b \in P$.

The T–transform For the DC d, define Td by the rule

$$Td(h) = \{(a,b) : d(a,b) \leq h\},$$

noting that $Td(h)$ is a reflexive symmetric relation. $Td(h)$ is an equivalence relation for all h if and only if d is an ultrametric. When ordered by set inclusion, the smallest reflexive symmetric relation is denoted $R_\emptyset$, and is defined by $R_\emptyset = \{(a,a) : a \in P\}$, and the largest one is given by $P \times P$. It is easy to show that the reflexive symmetric relations then form a Boolean algebra isomorphic to the power set of the two element subsets of P.

Relations of the form $Td(h)$ are called *threshold relations* of d, and the *proper* threshold relations are those other than $R_\emptyset$.

There is a natural well known bijection between ultrametrics and indexed nested sequences of equivalence relations, the top one being $P \times P$.

A cluster method is then a mapping $d \mapsto F(d)$ where d and F(d) are DCs. The usual algorithm takes $F(d)$ to be an ultrametric.

If $|P| = p$, and $k = p(p-1)/2$, then DCs may be viewed as vectors in the positive cone of a k-dimensional Euclidean vector space, and cluster methods may be viewed as mappings on this positive cone. Any of the usual metrics for Euclidean spaces may then be used. In particular, we use Δ_0 which is defined by

$$\Delta_0(d_1, d_2) = \max\{d_1(a,b) - d_2(a,b)| : a, b \in P\},$$

and is based on the L_∞-norm. Continuity, left continuity, and right continuity of a cluster method then all have their expected meanings.

It is easy to justify continuity as a desirable condition for a cluster method. The input data may very well have small errors, and it would be nice if a small error for the input would translate to a small error for the output. But in their book Mathematical Taxonomy, N. Jardine and R. Sibson showed that in the presence of continuity and certain other properties, the only acceptable cluster method is single-linkage clustering. This is defined by taking $Td(h) = \gamma \circ Td(h)$, where $\gamma(R)$ is the equivalence relation generated by the relation R.

Properties of Cluster Methods

We rephrase here some of the axioms that were originally introduced by Jardine and Sibson.

(JS1) *Idempotent* $F = F \circ F$.

(JS3) *Scale invariance.* $F(\alpha d) = \alpha F(d)$ for all $\alpha > 0$.

(JS3a) *Monotone equivariance* $F(\theta d) = \theta F(d)$ for every order automorphism θ of the nonnegative reals.

(JS5) *Isotone* $d_1 \leq d_2$ implies that $F(d_1) \leq F(d_2)$.

(JS5a) *0-isotone* $Td_1(0) = Td_2(0)$, then $d_1 \leq d_2$ implies $F(d_1) \leq F(d_2)$.

Theorem: For a monotone equivariant cluster method F, the following conditions are equivalent:

1. There exists a mapping γ on the reflexive symmetric relations such that for every DC d, $TF(d) = \gamma \circ Td$.
2. F is continuous.
3. F is right continuous.

Theorem Let F be monotone equivariant. Then F is left continuous if and only if there is a family $(\gamma_R)_{R \in \Sigma(P)}$ of mapping on $\Sigma(P)$ such that $TF(d) = \gamma_{Td(0)} \circ Td$.

F being isotone has unexpected consequences.

Theorem If the image of F contains all ultrametrics, and if F satisfies JS1 and JS5, then $F(d) \leq d$ for every DC d.

Theorem: If F satisfies JS3 and JS5, then F is left continuous. It is in fact continuous at all definite DCs.

Here is an example illustrating this Theorem. Take $F(d) = \underline{0}$ if d is not definite, and $F(d)$ to be single linkage clustering on the definite DCs.

Theorem Let F be monotone equivariant. Then JS5a is equivalent to left continuity, and JS5 is equivalent to continuity.

Thus continuity plus monotone eqivariance rules out almost all cluster algorithms that are commonly used by investigators. We will argue that the important property of continuity is ordinal in nature rather than metric.

Clustering Data Having Ordinal Significance

A DC d has ordinal significance if the values of d have no meaning, only whether one of $d(a,b) < d(x,y)$, $d(a,b) > d(x,y)$ or $d(a,b) = d(x,y)$ is true. In their book, Jardine and Sibson argue that one should use a *monotone equivariant* cluster method. This is a cluster method F having the property that $F(\theta d) = \theta F(d)$ for every DC d, and every order automorphism θ of $\Re_0^+$. This is a rather strong assumption, and in a later paper Sibson argues that it suffices to use a cluster algorithm that preserves *global order equivalence*, which is denoted $\sim_g$, and defined by the rule that $d_1 \sim_g d_2$ if and only if there is an order automorphism θ of $\Re_0^+$ such that $d_1 = \theta \circ d_2$. Thus one wants $d_1 \sim_g d_2$ to imply that $F(d_1) \sim_g F(d_2)$. Two cluster methods F, G are globally order equivalent if $F(d) \sim_g G(d)$ for every DC d defined on P. It turns out that every cluster method F that

> preserves global order equivalence and has the property that
> the image of $F(d)$ cannot have more members than the image of d

is globally order equivalent to a monotone equivariant cluster method, so we have not moved far from monotone equivariance.

But let $P = \{a, b, c\}$ with $d_1(a,b) = 0, d_1(a,c) = 1$ and $d_1(b,c) = 3$. If $d_2 = d_1 + 1$, then d_1 and d_2 are not globally order equivalent; yet they are equivalent in a way that we need to preserve. The proper definition is to say that d_1 and d_2 are weakly order equivalent (denoted $d_1 \sim_w d_2$) in case $d_1(a,b) < d_1(x,y) \iff d_2(a,b) < d_2(x,y)$. But now things are not so nice. A monotone equivariant cluster method need not preserve weak order equivalence. One can characterize when a cluster method that preserves weak order equivalence is weakly order equivalent (obvious definition) to a monotone equivariant cluster method.

The big question now is this. What in the world does any of this have to do with continuity in the Δ_0 metric? Hang on. A clue is coming.

The Connection with Continuity

If continuity is a desirable condition, it would be very nice to find a continuous cluster method other than single linkage clustering. Where does one look? Let's start by seeing if there is anything that all continuous cluster methods might have in common.

For any DC d, define the *mesh width* of d by

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

where the image of d is $0 = h_0 < h_1 < \cdots < h_t$.

Fundamental Result: If $\Delta_0(d, d') < \mu(d)$, then

$$d(a,b) < d(x,y) \Longrightarrow d'(a,b) < d'(x,y).$$

Use $d \preceq d'$ to denote the fact that $d(a,b) < d(x,y) \Longrightarrow d'(a,b) < d'(x,y)$. Note that $d \sim_w d' \Longleftrightarrow d \preceq d'$ and $d' \preceq d$. So suddenly there is a connection between metric properties of Δ_0 and ordinal considerations. Indeed, if $d_n \to d$, there must exist a positive integer N such that $n \geq N \Longrightarrow d_n \preceq d$. There is a weak converse connection given by the fact that $d \preceq d'$ implies the existence of d'' such that $d' \sim_w d''$ and $\Delta_0(d, d'') < \mu(d)$. In fact $d \preceq d'$ is equivalent to d being arbitrarily close to some d'' with d'' weakly order equivalent to d'.

Theorem: $d \preceq d'$ if and only if there is a sequence (d_n) of DCs all weakly order equivalent to d' such that $d_n \to d$,

Theorem: $d \preceq d'$ if and only if every proper splitting relation of d is a splitting relation of d'.

Definition. A cluster method F is *ordinally continuous* if $d \preceq d' \Longrightarrow F(d) \preceq F(d')$.

It is natural to conjecture that monotone eqivariance together with ordinal continuity might imply continuity. Here is an example is given showing this to be false. Let $R_1, R_2, \ldots, R_n$ denote the proper splitting relations of d. Take as the splitting relations for $F(d)$ those R_i that happen to be equivalence relations. Assign each such equivalence relation the level at which it came into being for d. This cluster method is monotone equivariant, order continuous, but not continuous. We illustrate this concretely.

Let $P = \{a, b, c\}$, and define $d(a,b) = d(a,c) = 1$, with $d(b,c) = 2$. d' is defined by $d'(a,b) = 1, d'(a,c) = 1 + \varepsilon, d'(b,c) = 2$, where $0 < \varepsilon < 1/4$. Note that $\mu(d) = 1/2$, and $\Delta_0(d, d') < 1/4$. The reader can verify that $P \times P$ is the only proper splitting relation of $F(d)$, while $F(d')$ has $P \times P$, as well as $R_\emptyset \cup \{(a,b),(b,a)\}$. It follows that $Fd(a,b) = Fd(a,c) = Fd(b,c) = 2$, while $Fd'(a,b) = 1$ with $Fd'(a,c) = Fd'(b,c) = 2$. Thus $\Delta_0(d, d') = \varepsilon$, while $\Delta_0(Fd, Fd') = 2$. Letting $\varepsilon \to 0$, it follows that F is not continuous.

If we take the view that it is only the partitions that $F(d)$ produces that are of interest, and not the levels at which they occur, then if we define a cluster method G to be single linkage clustering with the levels of the output rank ordered, then G is just as good as single linkage as a cluster algorithm. Thus we want conditions of a cluster method that tell us when the method is weakly order equivalent to a continuous cluster method. The only clear fact for such a cluster method is that it

must be order continuous. Such a cluster method need not be isotone, nor need it preserve multiplication by a positive scalar α.

Fundamental Question: Find necessary and sufficient conditions on a cluster method F so that $F$is weakly order equivalent to a continuous cluster method.

Examples are wanted (if there are any) of useful continuous cluster methods other than single linkage clustering.

Is continuity the issue? Complete linkage clustering is not continuous, but does have the property that $d \sim_w d' \Longrightarrow F(d) \sim_w F(d')$. If we restrict ourselves to DCs having no tied values, then $\Delta_0(d, d') < \mu(d) \Longrightarrow d \sim_w d'$. Is this the key property that needs to be preserved?