Classification in Mathematics, Discrete Metric Spaces, and Approximation by Trees

To Willem Kuyk

Michiel Hazewinkel
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
email: mich@cwi.nl

This is partly an introductory survey paper to clustering and classification problems with particular emphasis on the classification of lists of key words and phrases from a given scientific domain such as mathematics. In addition the paper contains a number of new concepts and results; a number of open questions, and some as yet untried embryo clustering ideas. New are the idea of Urysohn distance (Section 3), the idea of using Lipshitz distance (Section 4), the universal lower bound in terms of Lipshitz distance for any fixed depth hierarchical classification scheme (Section 8), the optimality of single link clustering with respect to Lipshitz distance (Section 8); in addition there are new results on what I have started to call the Buneman tree of a metric space (Section 9); also new are the ideas of third party support (Section 11) and power set metrics (Section 10).

1. The central mathematical problem

The central mathematical problem that I am concerned with in these notes is simply stated: given a discrete metric space, or more generally, a dissimilarity space (definition below), what is the 'best' classification tree to describe it. The motivation comes from classification problems in mathematics and other sciences and the intended application is to a large set of 'key words' and 'key phrases' of a discipline like all of mathematics, or all of physics, or substantial subfields like 'Lie algebras' or 'Surface physics'. It could help for much of what follows to keep this intended application in mind. In particular there is no claim that everything in this essay is relevant for all kinds of taxonomic problems.

It is a little peculiar and disturbing how pervasive and popular tree classification schemes are. One even finds statements in the scientific literature to the effect that if you cannot organize your knowledge in the form of a tree it is not worth doing anything. I must disagree with emphasis. Indeed, I do not think that our brains store knowledge in that way. I submit that the preoccupation with trees as an organizing principle comes from the tyranny of the classically

Vierde serie Deel 13 No. 3 november 1995, pp. 325–361 Nieuw Archief voor Wiskunde

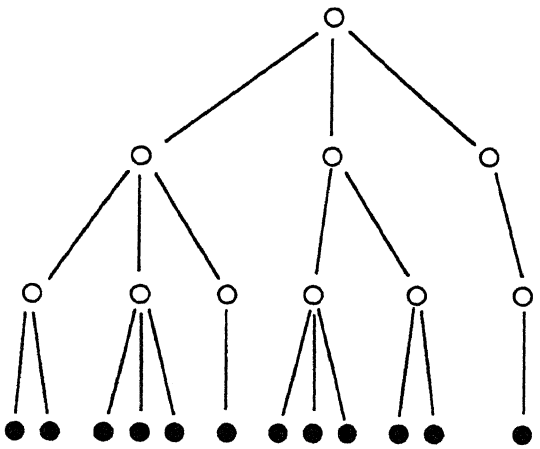


FIGURE 1.

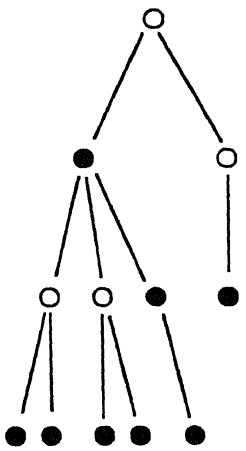


FIGURE 2.

printed word. It is just about the only kind of interconnectedness scheme that can be conveniently printed in book form.

To formulate the mathematical problem of classifying things precisely requires making precise what is meant with a classification tree and what is meant with 'best'. Also, what is a discrete metric space (for completeness) and a dissimilarity space.

1.1. Dissimilarity spaces

A *dissimilarity space* is a (finite) set M together with a function $d : M \times M \to \mathbb{R}$ such that

$$d(x,y) = d(y,x) \geq 0, \qquad \forall x, y \in M \tag{1.1.1}$$

$$d(x,x) = 0, \qquad \forall x \in M \tag{1.1.2}$$

A dissimilarity space M is a *metric space* if in addition:

$$d(x,y) = 0 \Rightarrow x = y \tag{1.1.3}$$

$$d(x,y) \leq d(x,z) + d(y,z), \qquad \forall x, y, z \in M \tag{1.1.4}$$

If (1.1.4), for distinct x, y, z, is always satisfied with the strict unequal sign, I shall say that the *strict triangle inequality* holds.

1.2. Classification trees

A (undirected) *graph* (without multiple edges) Γ is a finite set V, of which the elements are called nodes or vertices, together with a set E of unordered pairs

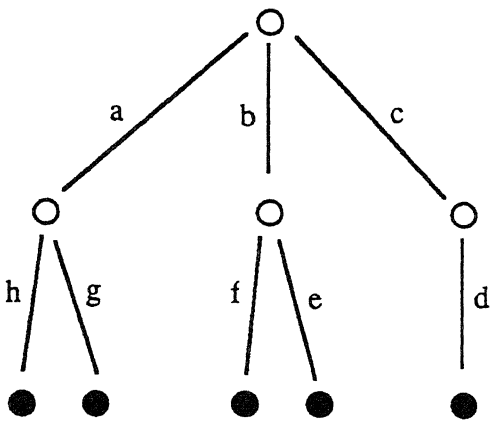


FIGURE 3.

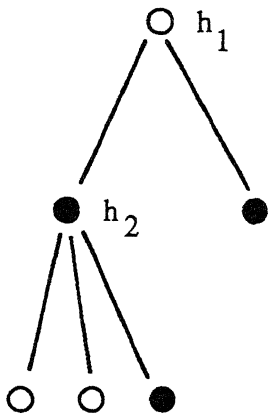


FIGURE 4.

of nodes called edges. We shall only consider graphs without loops, i.e. there are no edges of the form $\{x, x\}$. A path in Γ from a vertex x to a vertex y of k steps is a sequence of edges $e_1, \ldots, e_k$ so that $x \in e_1$, e_i and e_{i+1} have precisely one node in common, and $y \in e_k$. A node x is *incident with* an edge e iff $x \in e$. The *degree of a node* is the number of edges it is incident with.

A *tree* is a graph such that for every pair of vertices there is precisely one path from the one to the other. A *rooted tree* is a tree with one distinguished node called the root. The nodes which are maximal in number of steps from the root are called the leaves. These are also precisely the nodes of degree 1.

A *classification tree* for a dissimilarity space (M, d) is a rooted tree, Γ, together with a mapping of $\phi : M \to V$ to the nodes of Γ. For a *standard classification tree* for M, the mapping ϕ is required to be to the leaves of Γ. Often this mapping will be injective, but this need not necessarily be the case, especially when distances zero occur between distinct elements of the dissimilarity space (M, d).

That trees should be used for classification purposes is natural. Their role in clustering is emphasized in e.g. [23].

For a *standard classification tree*, the points of M are supposed to be the leaves of the rooted tree. For a nonstandard one, this need not be the case. In the illustrations I shall use small black filled circles for vertices of the tree that correspond to elements of M and small white circles for nodes that do not correspond to elements of M. Thus Figure 1 depicts a standard classification tree.

A classification tree is of *uniform depth* if the number of steps from the root to a leaf is always the same. Thus Figure 1 is a uniform depth standard classification tree of depth 3. Figure 2 depicts a nonstandard classification tree of non-uniform depth.

Let us further admit that the edges of the tree, or the nodes, can have

weights attached to them. The leaves of a tree, rooted or not, will always have weight zero; all other weights will be required to be strictly positive. Thus, above on the left, there is an edge-weighted standard classification tree of uniform depth 2 (Figure 3) and on the right there is a nonstandard and nonuniform node-weighted classification tree (Figure 4). The latter are special kinds of edge-weighted trees: the edges issuing from a node towards its children (viewed from the root) are all given half the weight of that node. By convention a non edge-weighted tree has all edges weighted by 1/2. An edge–weighted graph is also called a *network*; it is a *tree network* if the underlying graph is a tree. Every tree network, Γ, and more generally every network, defines a path length and from that a distance on its set of vertices by:

$$l(p) = \sum_{i=1}^{k} w(e_i), \quad p = e_1 \dots e_k \tag{1.2.1}$$

where p is a path in Γ, and $w(e)$ denotes the weight of an edge e, and

$$d_\Gamma(x, y) = \min_p \{l(p)\} \tag{1.2.2}$$

where p runs over all paths from x to y in Γ. Of course if Γ is a tree there is precisely one path from x to y. The distance on M defined by a standard or nonstandard, uniform or nonuniform, edge-weighted or not, classification tree $\phi : M \to \Gamma$ for (M, d) is a new dissimilarity d_ϕ on M defined by

$$d_\phi(x, y) = d_\Gamma(\phi(x), \phi(y)) \tag{1.2.3}$$

If the classification $\phi : M \to \Gamma$ is injective this will be a new metric on M.

1.3. Best approximating classification tree

Now, to define the notion of a best approximating classification tree of a given kind it remains to define a suitable notion of distance between different metrics or dissimilarity measures on a set, or more generally, between different metric spaces. A very easy choice would be to observe that metrics and dissimilarity measures on a given set M are simply special kinds of functions on $M \times M$ and to use any of the well known distance notions for functions on a set such as l_2-distance, giving:

$$\delta_{l_2}(d_1, d_2) = \left(\sum_{x,y \in M} (d_1(x, y) - d_2(x, y))^2 \right)^{1/2} \tag{1.3.1}$$

I shall use δ's to designate distances between (different) metric spaces, and d's to denote the distances defined on a given set.

Alternatively one could use the uniform, or l_p, or weighted- l_2, or any other favourite norm on function spaces. These are in fact the usual distances traditionally used in mathematical taxonomy, [18, 33, 34, 53]. There is something

very unsatisfactory in this. Distances are very special kinds of functions and to compare them using notions designed for dealing with the immeasurably larger collection of all functions is a little like using a classification scheme for all of the vegetable and animal kingdom to grade a batch of apples according to a handful of sizes.

Below I shall discuss a few other notions of distance between metric spaces which are more specific for that class of objects. Perhaps the most natural is Urysohn distance (Section 3), but it has the disadvantage that I know of no way to calculate it in concrete cases. Another which is very calculable and which seems to be well-suited to mathematical taxonomy is Lipshitz distance (Section 4). Finally in Section 5 a few distances between graphs are mentioned, which may be worth considering when dealing with classification problems.

Once a notion of distance between metric spaces (or dissimilarity spaces) has been chosen, the problem of finding the best approximating classification tree (of a specified type) is well defined, though, of course, there may be more than one solution.

2. The motivation

The classification scheme currently used in mathematics (The Mathematics Subject Classification Scheme (MSCS), developed by Mathematical Reviews and Zentralblatt für Mathematik and Grenzgebiete) is essentially top-down. It incorporates enormous expertise and a vast amount of effort from specialists in the various (sub)fields; it is very valuable as an information finding tool, but I am convinced that even better tools can be developed.

To a large extent the MSCS reflects how mathematics evolved historically and split into finer and finer subspecialisms. There is no check on whether the classification tree that is used reflects the actual, present day occurrence in the mathematical literature (in relation to each other) of the various technical mathematical terms used.

Exactly the same things can be said about PAC, the most used Physics and Astronomy classification scheme.

The mathematical literature currently grows at well over 50 000 journal and proceedings articles per year and with more than 1200 books a year. All in all, world-wide, even with the current funding crises, a considerable amount of (financial) resources is spent on new mathematical research; very little is spent on making sure that a result (a mathematical fact), once discovered, will ever be found again when needed, [26, 29]. The problem is serious and occasionally recognized as such, [13].

The MSCS is a uniform depth classification tree with near 5000 leaves. As a result whole books and (rarely) even whole journals can fall within a single leaf, and the number of articles that are assigned to a given leaf classification number can run in the thousands. Here is a small random sample with the minimum and maximum (in this sample) marked:

03E55 (Large cardinals):	880	
03F55 (Intuitionistic mathematics	575	
03G05 (Boolean algebras in logic)	807	
05C05 (Trees)	3147	
14D22 (Fine and course moduli spaces)	163	←
14H52 (Elliptic curves)	326	
14L05 (Formal groups)	465	
20F28 (Automorphism groups of groups)	398	
20G15 (Linear algebraic groups over arbitrary fields)	1254	
31B05 (Harmonic, sub- and superharmonic functions)	1081	
34D20 (Lyapunov stability for ODE)	4321	
35K22 (Evolution PDE of any order)	1453	
35L65 (Conservation laws for PDE of hyperbolic type)	1355	
46E15 (Banach spaces of .... functions)	1748	
46L35 (Classification of C*-algebras, factors)	636	
58F07 (Completely integrable systems)	1504	
90C10 (Integer programming)	4348	←

Inevitably there are overlaps. These are handled with 'see references'. All in all the number of such (indicated) overlaps (between nodes at the finer levels) is quite small.

2.1. Control list (glossary, thesaurus)
I estimate that a control list (also called thesaurus or glossary) is needed of some 120000 items (= key-phrases and key-words) in order to give reasonably effective descriptions of journal articles by assigning them items from this control list. Ideally, these items will also provide descriptions of the leaves of the MSCS in terms of (probably overlapping) clusters.

2.2. The web of mathematics
Once the control list is established, the items in it can be considered as the nodes of an edged weighted graph where two items are connected if they occur together in the same mathematical paper (or other suitable unit) and where the weight is inversely proportional to how many of such co-occurences there are. There are of course many ways of doing this. This topic of turning similarity indicators (such as co-occurences) into a metric or dissimilarity has had considerable attention in the classification and taxonomy literature, and I will say nothing more about it here.

There results a network of nodes and labelled links, a web, which I call the *web of mathematics*. Such a web will form one of the main navigation structures in the future hypertext–organized, interactive, electronic, CDROM–based, four-fold expanded *Encyclopaedia of Mathematics*, [24, 25, 27] (the KREEM project), and an electronic version of [5], the VEIG project.

One thing which will no doubt come out of this study is the discovery of numerous notions from different parts of mathematics which are completely or mostly the same. I was once lucky that way, [30], and have no doubt that there are many such cases, mostly unrecognized, except by an occasional individual.

One function of the control list is to provide the list of such alternatives when they exist. Whence the alternative name 'thesaurus' for the control list.

2.3. Local search
This web also provides the possibility for the creation of a refined mathematical information finding tool in that one can search for occurences of phrases and papers that are within a specified distance of a given node (local search).

2.4. BUC'M, bottom-up classification in mathematics
Once the web of mathematics is constructed one has a metric space, and one can try to find the best classification tree (of a desired kind), approximating it. This is the intended application of the research program outlined in this essay. I call it BUC'M, Bottom-up Classification in Mathematics. It will, I expect, provide valuable supplementary classification information to the top-down MSCS scheme, [28].

It works from the actual scientific data in the form of the published literature and aims to establish the 'best' classification tree on that basis and then to compare the results with the existing classification scheme.

Once established it will also provide tools for the automatic assignment of key-words and key-phrases to articles and the automatic classification of articles in terms of the MSCS and BUC'M classification trees. These matters, providing key-words and key-phrases, are currently often left to authors (a haphazard procedure), scientific editors (almost equally unreliable), or even desk-editors employed at scientific publishers (totally unreliable).

3. Urysohn distance

Probably the most intrinsic notion of distance between metric spaces is what I shall call here Urysohn distance. To define it I need the notion of the Hausdorff distance between two subsets of a metric space.

3.1. Hausdorff distance
Let X, Y be two subspaces of a metric space . Then the Hausdorff distance between them is defined by:

$$Hd(X,Y) = \max\{\sup_{x\in X}\inf_{y\in Y}\{d(x,y)\}, \sup_{y\in Y}\inf_{x\in X}\{d(x,y)\}\} \tag{3.1.1}$$

(Sometimes a variant is employed in which instead of the max, the sum of the two terms is taken; there are also of course all the possibilities of l_p-type).The Urysohn distance between two metric spaces is now defined as

$$\delta_U(M_1, M_2) = \inf_{\alpha,\beta}\{Hd(\alpha(M_1), \beta(M_2)\} \tag{3.1.2}$$

where the infimum is taken over all isometries α, β of M_1 and M_2 into a third metric space N.

3.2. Urysohn spaces
There exist universal metric spaces. More precisely URYSOHN, see [37, 56, 57, 58], proved that there is a unique complete metric space that is $\aleph_0$-universal and $\aleph_0$-homogeneous. Here these phrases mean the following. A metric space U is $\aleph_0$-universal if for every metric space M of countable cardinality, there is an isometric inbedding $M \to U$; a metric space U is $\aleph_0$-homogeneous if for every finite metric space Y and subspace X of Y and isometry $\alpha : X \to U$, there is an isometry $\beta : Y \to U$, that restricts to α on X. Actually such spaces exist for all cardinals and not only for $\aleph_0$. One way to construct them is as inductive limits over all (isometry classes of) relevant spaces with respect to the filtered system of isometries between them.

This means that in the definition (3.1.2) it is not necessary to vary N over all (finite) metric spaces; it suffices to consider only isometries of M_1 and M_2 into a Urysohn space U.

3.3. Theorem. The Urysohn distance as defined by (3.1.2) is a metric on the set of all isometry classes of finite metric spaces.

PROOF. It is obvious that $\delta_U(M_1, M_2) \geq 0$, and that $\delta_U(M_1, M_2) = 0$ if and only if M_1 and M_2 are isometric. It remains to prove the triangle inequality. So let $\alpha_1 : M_1 \to U'$, $\alpha_2 : M_2 \to U'$, and, $\beta_2 : M_2 \to U$, $\beta_3 : M_3 \to U$ be isometric inbeddings such that

$$Hd(\alpha_1(M_1), \alpha_2(M_2)) \leq \delta_U(M_1, M_2) + \varepsilon \tag{3.3.1}$$

$$Hd(\beta_2(M_2), \beta_3(M_3)) \leq \delta_U(M_2, M_3) + \varepsilon \tag{3.3.2}$$

for some positive (small) real number ε. Replacing, if necessary, U' with the union of the images of M_1 and M_2, we can assume that U' is finite and view, via α_2, M_2 as a subspace of U'. By the homogeneity property of U, there exists an isometry $\gamma : U' \to U$, that extends the isometry β_2. Then $\gamma\alpha_2 = \beta_2$, and $Hd(\alpha_1(M_1), \alpha_2(M_2)) = Hd(\gamma\alpha_1(M_1), \beta_2(M_2))$, and hence

$$\begin{aligned} &\delta_U(M_1, M_3) \leq Hd(\gamma\alpha_1(M_1), \beta_3(M_3)) \leq \\ &Hd(\gamma\alpha_1(M_1), \beta_3(M_2)) + Hd(\beta_2(M_2), \beta_3(M_3)) \leq \\ &\delta_U(M_1, M_2) + \delta_U(M_2, M_3) + 2\varepsilon \end{aligned} \tag{3.3.3}$$

This proves the triangle inequality for the Urysohn distance. □

The Urysohn distance as defined by (3.1.2) (with a fixed Urysohn space U if desired instead of all finite metric spaces N) is completely intrinsic. However, I know of no way to compute this distance between two given metric spaces.

These ideas extend to dissimilarity spaces. I.e. there should also exist a universal dissimilarity space for all finite dissimilarity spaces, and then that one can be used for defining a Urysohn dissimilarity between dissimilarity spaces, using the Hausdorff dissimilarity between two subsets of a dissimilarity. All

these will be defined completely analogously to (3.1.1) and 3.1.2).

4. LIPSHITZ DISTANCE

A completely different idea of comparing metric spaces is based on the idea of Lipshitz mappings. This is probably the origin of the name 'Lipshitz distance'. It is eminently calculable and has the additional merit of having proved its worth in another field of mathematics, viz. Riemannian geometry, [17, 20, 46]. It is essentially limited to comparing two different metrics on the same underlying set, or, equivalently, on two sets of the same cardinality. The paper [40] contains some results concerning inbeddings of finite metric spaces into Euclidean spaces with minimal Lipshitz distance between the original metric and the induced metric.

4.1. Lipshitz distance

Consider a set M and two metrics, d_1, d_2 defined on it. The *distortion* of d_2 with respect to d_1 is defined by

$$\text{distor}(d_2, d_1) = \sup \frac{d_2(x,y)}{d_1(x,y)} \tag{4.1.1}$$

where the sup is taken over all $x, y \in M$, $x \neq y$. The Lipshitz distance between d_1, d_2 is now defined as

$$\delta_L(d_1, d_2) = \log(\text{distor}(d_2, d_1)\text{distor}(d_1, d_2)) \tag{4.1.2}$$

Note that if the two distances are proportional, their Lipshitz distance is zero. This is really an advantage in our setting because in classification problems like these a constant scalar factor should not matter.

It is easy to see that:

4.2. Proposition. The Lipshitz distance δ_L defines a metric on isometry classes up to a scalar factor of metrics on a fixed set M.

5. VARIOUS GRAPH DISTANCES

Several distances between (not edge labelled) graphs have been defined in the literature. The underlying ideas can probably be usefully extended to the case of edge labelled graphs, i.e. networks. I mention a few. A selection from the literature is [1, 2, 8, 36, 52, 54, 61], and the material in this section comes from there (mostly [54]). I shall not do anything with these distances in this essay. All the same it will certainly be interesting to compare these distances with Urysohn distance and Lipshitz distance.

For two G, H graphs define:

$$U_m(G), \text{ the set of isomorphism classes of subgraphs of } G \text{ with } m \text{ vertices;} \tag{5.0.1}$$

$$U(G), \text{ the union of all the } U_m(G) \tag{5.0.2}$$

$u(H, g)$, the number of subgraphs of G isomorphic to H; (5.0.3)

$G \cap H$, the largest, in number of vertices, graph that is isomorphic to both a subgraph of G and to a subgraph of H; (5.0.4)

$G \cup H$, the smallest, in number of vertices, graph that contains a subgraph isomorphic to G and a subgraph isomorphic to H. (5.0.5)

5.1. Zelinka distance
The Zelinka distance between two graphs is now defined as

$$\delta_Z(G, H) = \max\{\#V(G), \#V(H)\} - \#V(G \cap H) \tag{5.1.1}$$

where $\#X$ denotes the number of elements of a set X.

5.2. Some other graph distances
Define

$$m_1(G, H) = \min\{m : U_m(G) \neq U_m(H)\} \tag{5.2.1}$$

$$m_2(G, H) = \min\{m : \exists L \text{ such that } \#V(L) = m, \; u(L, G) \neq u(L, H)\} \tag{5.2.2}$$

Using these, one defines the graph distances

$$\delta_1(G, H) = \begin{cases} \max\{\#V(G), \#V(H)\} + 1 - m_1(G, H) & \text{if } G \not\simeq H \\ 0 & \text{if } G \simeq H \end{cases} \tag{5.2.3}$$

$$\delta_2(G, H) = \begin{cases} \max\{\#V(G), \#V(H)\} + 1 - m_2(G, H) & \text{if } G \not\simeq H \\ 0 & \text{if } G \simeq H \end{cases} \tag{5.2.4}$$

$$\delta_3(G, H) = \begin{cases} m_1(G, H)^{-1} & \text{if } G \not\simeq H \\ 0 & \text{if } G \simeq H \end{cases} \tag{5.2.5}$$

$$\delta_4(G, H) = \begin{cases} m_2(G, H)^{-1} & \text{if } G \not\simeq H \\ 0 & \text{if } G \simeq H \end{cases} \tag{5.2.6}$$

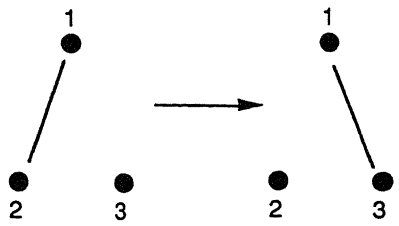
$$\delta_5(G, H) = \#E(G) + \#E(H) - 2\#E(G \cap H) + |\#V(G) - \#V(H)| \tag{5.2.7}$$

$$\delta_6(G, H) = -\#E(G) - \#E(H) + 2\#E(G \cap H) + |\#V(G) - \#V(H)| \tag{5.2.8}$$

If in (5.2.7) one replaces the three E's with V's, the Zelinka distance is re-obtained (more precisely twice the Zelinka distance).

5.3. Proposition. The formulas (5.1.1), (5.2.3)-(5.2.8) all define distances on the set of isomorphism classes of graphs.

Still another distance is the edge *rotation* distance defined on the set of (isomorphism classes of) graphs with a fixed number of edges and vertices. That distance is measured by the least number of edges that must be rotated to get from one graph to the other. Here an edge rotation is a local modification operation such as illustrated on the right. Here on the left side of the picture there is not supposed to be an edge from 1 to 3. It is a not totally obvious fact that one can always go from one graph to any other (with the same number of edges and vertices) by means of such edge rotation operations.



6. CLUSTERING

A nonoverlapping clustering on a set of objects M is simply a standard classification tree for M. All clusterings in this section will be nonoverlapping. Overlapping clusterings will be discussed below in Section 10. To carry out a clustering procedure is to construct such a classification tree. More precise definitions follow.

6.1. Definitions

A single layer clustering on a set of objects M is simply a partion of M; i.e. a splitting up of M into clusters $C_1, C_2, \ldots, C_k$ such that

$$C_i \cap C_j = \emptyset \text{ if } i \neq j, \text{ and } \bigcup_{i=1}^{k} C_i = M \tag{6.1.1}$$

A (hierarchical) clustering of M is a collection $\mathcal{H}$ of subsets of M, such that:

$$\begin{aligned} &\{m\} \in \mathcal{H}, \ \forall m \in M \\ &M \in \mathcal{H} \\ &\forall C, D \in \mathcal{H}, \ C \cap D = \emptyset, \text{ or } C \subset D, \text{ or } D \subset C \end{aligned} \tag{6.1.2}$$

By adding to the sets making up a single layer clustering the whole set and the singleton sets (in so far as not already present) such a partition defines a hierarchical clustering of depth 2.

The classification tree of a hierarchical clustering $\mathcal{H}$ consists of the elements of $\mathcal{H}$, ordered by inclusion, with M as the root and the singleton sets as the

leaves.

There are a great many methods in the classification literature for obtaining clusterings. A few will be briefly mentioned below.

6.2. Single link clustering

Define the single link dissimilarity between two subsets X, Y of a metric space M as

$$\mathrm{sld}(X,Y) = \inf_{x\in X, y\in Y}\{d(x,y)\} \tag{6.2.1}$$

Then an algorithm for single link clustering can be described as follows:

1) Let $\mathcal{H}$ consists of all singleton sets. Set i=0. (6.2.2.)
2) Set $i := i + 1$. Find the minimal single link distance $\neq 0$ between the (current) elements of $\mathcal{H}$. Let d be this minimals ingle link distance. Let $\mathcal{T}$ denote a set of temporary clusters. Set $\mathcal{T} = \emptyset$.
3) Set $\mathcal{H} = \mathcal{H} \cup \mathcal{T}$. Joint any two clusters in $\mathcal{H}$ at sld d and add the new cluster to $\mathcal{T}$
4) Repeat 3 until no more clusters are left in $\mathcal{H}$ at sld d. Now let (the new) $\mathcal{H}$ be the union of (the old) $\mathcal{H}$ and the maximal elements of $\mathcal{T}$.
5) If $\mathcal{M} \in \mathcal{H}$, stop; otherwise go to step 2.

There is choice involved in step 3). But it is easy to see that this does not matter for the end result. Step 3) and 4) together can be conceptually cleaner described as follows.

Two clusters $\mathcal{H}$ in are related if their sld is equal to d. Take the transitive closure of this relation. From the unions of the elements in the equivalence classes for this equivalence relation and add those to $\mathcal{H}$. (6.2.3)

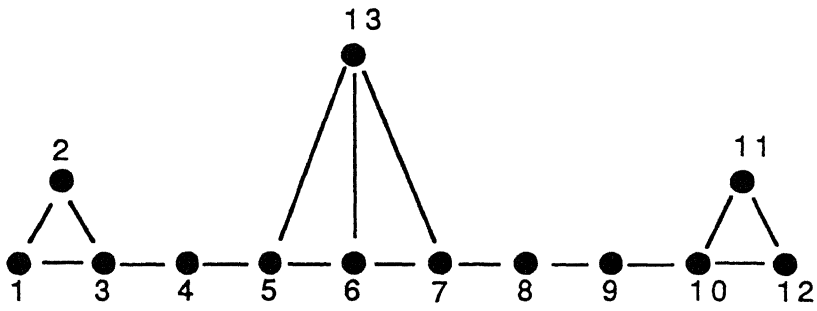


FIGURE 5.

In the network example of a metric space above all distances indicated by short edges are equal to 1, and the three long ones, i.e. the three distances to object

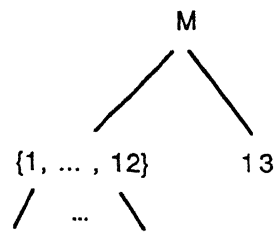


FIGURE 6.

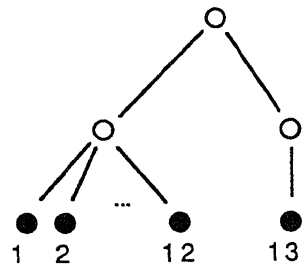


FIGURE 7.

13, are equal to 2. When single link clustering is applied to this example the result is as sketched in Figure 6.

This certainly illustrates what is perhaps the major defect of single link clustering, known as chaining: long extended chains of objects are put into single clusters. Other clustering methods have been invented to avoid this, but the price for that appears high. More about that below. More about single link clustering and its many good properties can be Sections 7 and 8.

A third way to think about single link clustering is as follows. List the distances that occur in the metric space under consideration.

$$0 < d_1 < d_2 < \ldots d_n \tag{6.2.5}$$

Given this here is a third description of single link clustering.

> For each $i = 1, \ldots, n$ draw the graph with as vertices the elements of M and with two elements connected if and only if their distance is less or equal to d_i. The connected components of these graphs for $i = 1, \ldots, n$ are precisely the members of the single link hierarchy for M. (6.2.6)

It is in fact more customary to use single link clustering to construct a fixed depth standard classification tree. That means that if in the procedure (6.2.6) for different i some clusters are the same these are retained and placed at all levels at which they occur. Thus the result of single link clustering on the network of Figure 5 will become as in Figure 7 above on the right rather than as in Figure 6.

6.3. Complete link clustering and some others

Complete link clustering as it is often described proceeds exactly like (6.2.2) except that the distance between two clusters is calculated differently. Instead of the dissimilarity 'dsl', the complete link distance 'cld' is used as defined by:

$$\mathrm{cld}(X, Y) = \sup_{x \in X, y \in Y} \{d(x, y)\} \tag{6.3.1}$$

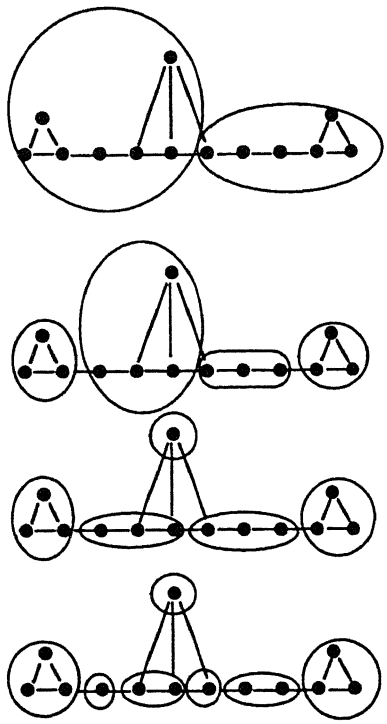


FIGURE 8.

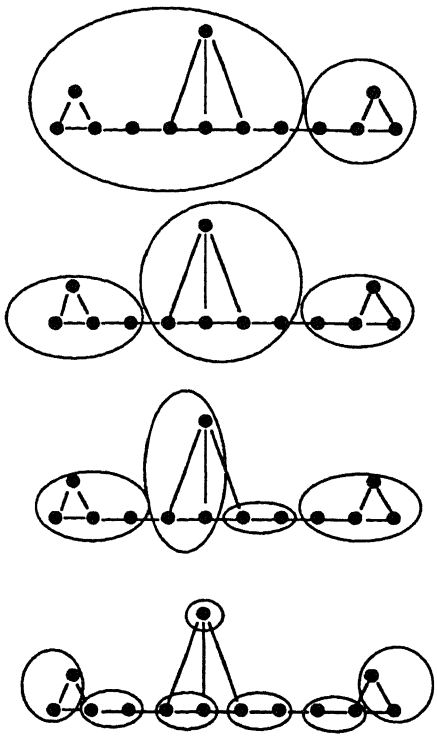


FIGURE 9.

Incidentally, cld is a distance; but sld is not (it does not necessarily satisfy the triangle inequality).

For complete link clustering the choices involved in step 3 of (6.2.2) do matter and the results obtained from different choices can be dramatically different. Two examples are above. It is clear that the results can be very unsatisfactory.

For this reason JARDINE and SIBSON [34] strongly argue that at each stage the transitive closure of the nearest distance relation between clusters should be taken. This makes complete link clustering exactly analogous to single link clustering in the form (6.2.3), except that the distances between clusters are measured by cld instead of sld. In the example above this version of complete link clustering gives exactly the same as single link clustering. The example is not large or diverse enough for the different ways of measuring distances between non-singleton sets to show up.

Still other ways, intermediate between complete link and single link clustering, have been frequently explored in the literature. For instance average link clustering, where the distance between two clusters is defined by

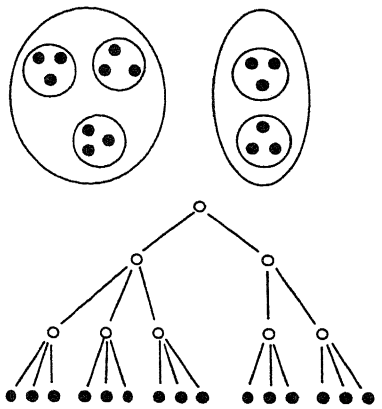


FIGURE 10.

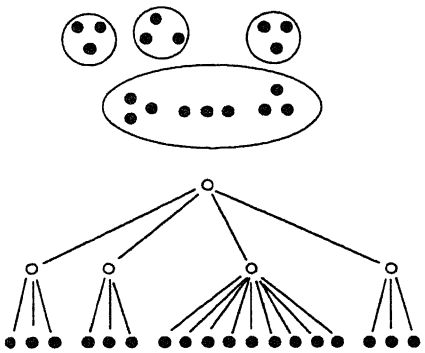


FIGURE 11.

$$\mathrm{ald}(X,Y) = (\#X)^{-1}(\#Y)^{-1} \sum_{x\in X, y\in Y} d(x,y) \tag{6.3.2}$$

In case the data set M is inbedded in a Euclidean space, still other ways of defining distances between sets can be natural leading to for instance the centroid clustering method. Hausdorff distance between sets, cf (3.1.1) above, could also be used. That possibility I have not (yet) seen in the literature. All of these suffer from the same choice difficulty as complete link clustering.

One can examine the many many clustering methods that are in the literature and that are being used from the axiomatic viewpoint. This means that one makes a list of desirable properties that a method should have and checks which methods satisfy which desirability axiom. This has been done for a substantial collection of methods in [16, 34, 58].

According to [34] only the single link method survives this critique except for overlapping variants of the single link method in which overlap between clusters is permitted of a maximal fixed in advance number of elements or a fixed in advance percentage of the size of the clusters involved.

Single link clustering has in addition a number of very nice properties (besides being also easy to implement). Some of these will be discussed in the next section.

7. SINGLE LINK CLUSTERING

As has already been mentioned a defect of single link clustering is 'chaining': quite distant groups of objects get joined by long tenuous chains. This can take quite serious forms as the example of Figure 10 above illustrates. On the other hand single link clustering has very many good properties including some categorical ones. It solves a universal problem in the sense of category theory. Most of this section is devoted to that.

7.1. Chaining example
Imagine that in a scientific subject classification situation like the one we are envisaging a quite nice constant depth standard classification tree has been found which really describes the interrelatedness of the concepts involved perfectly. As depicted in Figure 10. Now some talented maniac discovers a first tenuous connection between some quite distant parts. The intermediate concepts he defined for this purpose are interesting and taken up by a few more people. And shortly afterwards the situation is like depicted in the upper part of Figure 11, where the new objects are in grey. Now apply single link clustering to the new situation. Not only do to quite distant clusters get joined, but also a number of higher level groupings are destroyed (and the new classification tree in this example suddenly has depth two instead of three).

7.2. Ultrametric spaces
A metric d on a set M is ultrametric if it satisfies the following ultrametric property

$$d(x,y) \leq \max\{d(x,z), d(y,z)\}, \qquad \forall x,y,z \in M \tag{7.2.1}$$

It is easily checked that the metric defined on its set of leaves by a standard constant depth classification tree is always ultrametric. The inverse is also true.

7.3. Theorem (JOHNSON, [35])
A dissimilarity space (M,d) is an ultrametric space if and only if it is the metric defined by a standard, constant depth, vertex weighted classification tree for (M,d).

7.4. Subdominance of single link clustering
Let (M,d) be a metric space. Let $0 < d_1 < d_2 < \ldots < d_k$ be the distances occurring. Construct the single clink clustering constant depth standard classification tree by, say, (6.2.6). Give the vertices at i steps from a leaf the weight $\frac{1}{2}(d_i - d_{i-1})$, where $d_0 = 0$. The weighted classification tree thus defined defines a new metric on M, which will be denoted d_{sl}. Then:

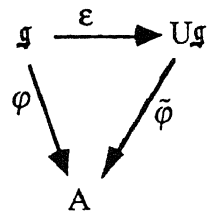
7.4.1. Theorem (see e.g. [34]). *The metric d_{sl} is uniformly smaller than d; i.e. $d_{sl}(x,y) \leq d(x,y)$ for all $x,y \in M$. The metric d_{sl} is ultrametric and for every ultrametric that is uniformly smaller than d, denoted $d' \leq d$, $d' \leq d_{sl}$. If d is ultrametric $d_{sl} = d$.*

Thus d_{sl} is the largest amoung the ultrametrics dominated by d. This property is referred to as the subdominance property of d_{sl}.

7.5. Universality of single link clustering
A property like the subdominance property above smells like universality properties and adjoint functor properties. An example of these, familiar to algebraists, is the universal enveloping algebra of a Lie algebra $\mathfrak{g}$, which is defined by the following universality property. It is an associative algebra $U\mathfrak{g}$, together

with a map of Lie $\varepsilon : \mathfrak{g} \to U\mathfrak{g}$ algebras such that for every map of Lie algebras from $\mathfrak{g}$ to an associative algebra A, $\phi : \mathfrak{g} \to A$, there is a unique map of associative algebras $\tilde{\phi} : U\mathfrak{g} \to A$, such that $\tilde{\phi} \circ \varepsilon = \phi$. In the present setting of metric spaces it is not so clear what is the appropriate notion of morphism which should be used. Let us for the moment settle on the category of metric spaces and contracting maps.

Within this setting, the single link classification space has the exact analogous property. Let M_{sl} be the set M with the metric given by its single link classification tree (which is ultrametric). The canonical map $M \to M_{sl}$ is contracting, and for every contracting map of M into an ultrametric space there is a unique factorization through M_{sl}. This is immediate from theorem 7.4.1. (The map is uniquely defined by the fact that as sets M and M_{sl} are identical, and the contractiveness of the factorization map then follows from the subdominace property). So, technically, the single link classification tree functor from metric spaces to ultrametric spaces is adjoint to the forgetful functor from ultrametric spaces to metric spaces.



In [51], ROUX gives an algorithm for turning a given metric space into an ultrametric one. It will be interesting to find out the relevance of this procedure to the single link way of producing an ultrametric from a given one.

8. LIPSHITZ DISTANCE AND CLUSTERING

8.1. Lipshitz distance and traditional clustering desirabilities

Consider for the moment the problem of finding a single layer clustering for a given discrete metric space and consider what trying to find minimal Lipshitz distance between d and the resulting classification tree distance implies. Let $M = C_1 \cup \ldots \cup C_m$ be a partition of M. Then the corresponding classification distance is

$$d_T(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } \exists i \text{ such that } x, y \in C_i \\ 2 & \text{otherwise} \end{cases} \tag{8.1.1}$$

By the definition of Lipshitz distance, see (4.1.2), we must try to minimize

$$\sup \frac{d(x,y)}{d_T(x,y)} \sup \frac{d_T(x,y)}{d(x,y)} \tag{8.1.2}$$

over all $x \neq y$. For subsets X and Y of M, let

$$s(X) = \inf_{x,y \in X, x \neq y} \{d(x,y)\}, \ \operatorname{diam}(X) = \sup_{x,y \in X} \{d(x,y)\} \tag{8.1.3}$$

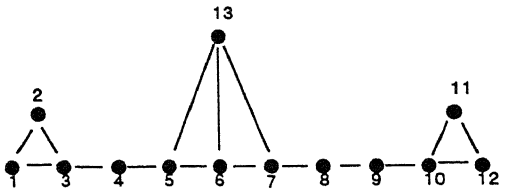


FIGURE 12.1.

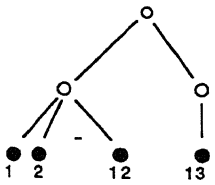


FIGURE 12.2.

and let $\mathrm{cld}(X,Y)$, $\mathrm{sld}(X,Y)$ be the complete link and single link distances between X and Y as defined in Section 6 above. Then minimizing the Lipshitz distance between d and d_T means trying to minimize simultaneously the following quantities

$$\frac{\mathrm{cld}(C_i,C_j)}{\mathrm{sld}(C_k,C_i)}, \quad \frac{\mathrm{cld}(C_i,C_j)}{2s(C_k)}, \quad \frac{\mathrm{diam}(C_i)}{s(C_k)}, \quad \frac{2\mathrm{diam}(C_i)}{\mathrm{sld}(C_k,C_i)}$$

There are no absolute relations between these four types of quantities. But something can be said 'generically'. As a rule not every pair of objects that are maximally close together should finish up in different clusters. That makes

$$\inf_k\{s(C_k)\} = s(X),$$

so that keeping the third type of expression in (8.1.4) under control means keeping the diameters uniformly reasonably small. (If all closest pairs are broken up, the fourth term will dominate the third and tend to get rather large).

Given that, the fourth type of term says to keep the clusters well separated and the first and second type of term then also argue in favour of keeping the diameters small and to keep the number of clusters relatively small as otherwise some of the cld distances will get near to $\mathrm{diam}(X)$. These are all considerations that are frequently found in the clustering literature and which are all of course intuitively right. Thus optimization with respect to Lipshitz distance seems to capture much of what seems intuitively right when looking for a good clustering.

Similar heuristic arguments give information on the sort of thing to try when trying to find the clusterings at a given level in a hierarchy.

8.2. Some examples

Let us again consider the example of Figure 5 and compare in terms of Lipshitz distance some intuitive clusterings with the results of single link clustering. Here are the clusterings that we shall examine. For convience the metric space itself is also reproduced. Give all vertices weight 1/2. The clusterings 12.2 and 12.3 come from single link clustering. The Lipshitz distances from the original metric to the new ones defined by these clusterings are as follows:

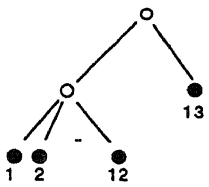


FIGURE 12.3.

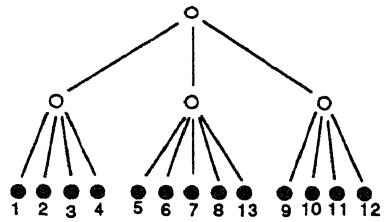


FIGURE 12.4.

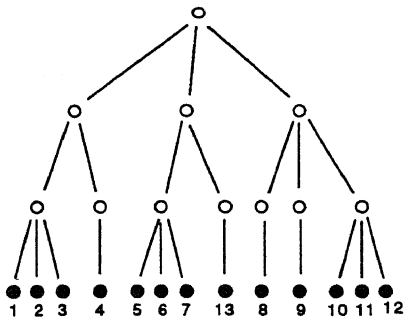


FIGURE 12.5.

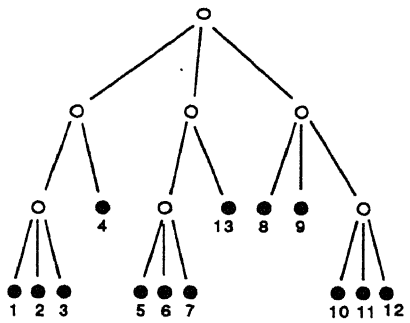


FIGURE 12.6.

12.2: log(9)
12.3: log(9)
12.4: log(9)
12.5: log(9)
12.6: log(15/2)
12.7: log(6)

Clusterings 12.4 and 12.5, though intuitively perhaps more appealing, are no better, in terms of Lipshitz distance, than single link clustering, showing that

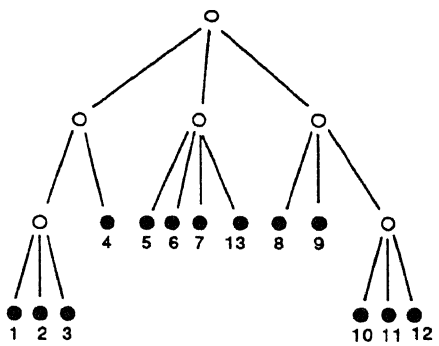


FIGURE 12.7.

it is in fact not so easy to do better. The clusters themselves are indeed better, but the price of breaking up closest neighbour pairs like 4, 5 or 7, 8 or 8, 9 for these clusterings is high. This is a general phenomenon which argues for single link clusterings. It will be made more precise below.

8.3. Step across separation and a universal lower bound
A path P from x to y in M is simply a sequence of points $x = x_0, x_1, \ldots, x_n = y$. The step length of P is equal to

$$sl(P) = \max_{i=0,\ldots,n-1}\{d(x_i, x_{i+1})\} \tag{8.3.1}$$

The step-across-separation of two unequal points $x, y \in M$ is defined as

$$sas(x, y) = \min_P\{sl(P)\} \tag{8.3.2}$$

where the minimum is taken over all paths P from x to y. Finally the 'distance step-across-separation quotient' of M is defined as

$$dsq(M) = \max_{x \neq y} \frac{d(x, y)}{sas(x, y)} \tag{8.3.3}$$

8.3.4. Theorem. Let $\pi = (\pi_0, \pi_1, \ldots \pi_n)$ be a constant depth standard classification tree on M with associated distance d_T. Then $\delta_L(d, d_T) \geq \log(dsq(M))$.

PROOF. Let $x, y \in M$ be such that the maximum in (8.3.3) is assumed for this pair. Moreover let $x = x_0, x_1, \ldots, x_n = y$ be a path for which the step length is equal to $sas(x, y)$. Let $d_T(x, y) = d_j$. Then

$$\text{distor}(d, d_T) \geq \frac{d(x, y)}{d_j} \tag{8.3.5}$$

Now consider the chain $x = x_0, x_1, \ldots, x_n = y$. Let $\pi_j = \{C_1, \ldots, C_m\}$. We can assume that $x, y \in C_1$. Suppose that in $\pi_{j-1}$the set C_1 splits up into the disjoint sets $C_{11}, \ldots, C_{1k}$. We can assume that $x \in C_{11}$, and then $y \notin C_{1l}$. Therefore there is a first x_i that is not in C_{11}. There are two possibilities. If $x_i \in C_1$, then $d_T(x_{i-1}, x_i) = d_j$; if $x_i \notin C_1$, then the finest partition with a set in it that contains both x_{i-1}, x_i has index $> j$, so that

$$d_T(x_{i-1}, x_i) \geq d_{j+1} > d_j \tag{8.3.6}$$

Also $d(x_{i-1}, x_i) \leq sas(x, y)$. Combining these two we see that

$$\text{distor}(d_T, d) \geq \frac{d_j}{sas(x, y)} \tag{8.3.7}$$

Inequalities (8.3.5) and (8.3.7) together prove the theorem. □

8.4. Single link clustering and Lipshitz distance

It is in fact not an accident that in the examples of Figures 12, all the constant depth clusterings did not improve on single link clusterings. We shall see here that in fact single link clustering is optimal (though there may be other clusterings that are equally Lipshitz close to the original metric).

Single link clustering yields the following hierarchy. Let $d_1 < d_2 < \ldots d_m$ be the distances that actually occur. Let G_j be the graph on M which has two vertices linked if and only if their distance in M is $\leq d_j$. Then the partition π_j has level d_j and it consists of the connected components of the graph G_j. Let d_{sl}, the single link clustering hierarchy distance, be the distance defined by this classification tree.

8.4.1. Theorem. Single link clustering is optimal with respect to Lipshitz distance. More precisely:

$$\delta_L(d_{sl}, d) = \log(dsq(M)) \tag{8.4.2}$$

PROOF. In view of theorem 3.4, it suffices to show that

$$\delta_L(d_{sl}, d) \leq \log(dsq(M)) \tag{8.4.3}$$

Let $x, y \in M$, $x \neq y$. Then $sas(x, y) = d_{sl}(x, y)$. It follows that

$$\text{distor}(d, d_{sl}) = \sup \frac{d(x, y)}{d_{sl}(x, y)} \leq dsq(M) \tag{8.4.4}$$

On the other hand, as has already been observed, theorem 7.4.1, $d_{sl}(x, y) \leq d(x, y)$ for all $x, y \in M$. Hence

$$\text{distor}(d_{sl}, d) \leq 1. \tag{8.4.5}$$

The combination of (8.4.4) and (8.4.5) establishes (8.4.3) and hence (8.4.2) and the theorem. □

9. THE BUNEMAN TREE OF A DISSIMILARITY SPACE

9.1. Tree metrics and the four-point condition

Let T be a tree and M a subset of the vertices of T. Let the edges of T be labelled with strictly positive numbers. This makes T a network (of a special kind), and induces on M a metric as follows:

$$d(x, y) = \text{length of the unique path in } T \text{ from } x \text{ to } y. \tag{9.1.1}$$

Such a metric on M will be called a tree metric. There is something special about this kind of metrics. They satisfy the socalled *four-point condition*:

$$d(x, y) + d(u, v) \leq \max\{d(x, u) + d(y, v), d(x, v) + d(y, u)\} \tag{9.1.2}$$

for all $x, y, u, v \in M$. This condition is easiest remembered by noting that

there are three ways of splitting up four points into two pairs. The sum of the distances of each such pair now must be smaller than the maximum of the other two sums. It follows that at least two of these sums are equal and maximal. Taking $u = v$, one sees that the four-point condition implies the triangle condition. The fourpoint condition in fact characterizes tree metrics.

9.1.3. Theorem. A metric on M is a tree metric if and only if it satisfies the four-point condition.

This was proved independently by BUNEMAN [6] and DOBSON [15]. Another proof is in [49]. Buneman, however, did much more. He associated in a canonical way, with every metric space M an edge labelled tree BT_M, together with a map $\beta : M \to BT_M$, such that if the metric on M was a tree metric than β is injective and the induced tree metric on M by is the original one. The next subsection summarizes the constructions and results from [6].

9.2. The Buneman construction
Let X be any set. A splitting, σ, of X is a two element partition of X, i.e.

$$\sigma = \{A, B\},\ X = A \cup B,\ a \cap B = \emptyset \tag{9.2.1}$$

Two splittings, $\sigma_1 = \{A_1, B_1\}$, $\sigma_2 = \{A_2, B_2\}$, are compatible if at least one of the four intersections

$$A_1 \cap A_2,\ A_1 \cap B_2,\ B_1 \cap A_2,\ B_1 \cap B_2 \tag{9.2.2}$$

is empty. Let σ_0 be the trivial splitting: $\sigma_0 = \{X, \emptyset\}$. Let $\Sigma = \{\sigma_j, \ldots, \sigma_m\}$ be a compatible set of splittings. Such a set of splittings defines a tree, T_Σ, as follows. The vertices of the tree are collections of sets

$$\{C_1, C_2, \ldots, C_m\},\ C_j \in \sigma_j,\ j = 1, \ldots, m \tag{9.2.3}$$

such that

$$C_i \cap C_j \neq \emptyset,\ \forall i, j = 1, \ldots, m \tag{9.2.4}$$

It turns out that there are precisely $m + 1$ vertices if $\sigma_o \neq \Sigma$, and m vertices otherwise. Two vertices

$$\{C_1, \ldots, C_m\},\ \{D_1, \ldots, D_m\} \tag{9.2.5}$$

are linked in the tree T_Σ if and only if there is precisely one index $j \in \{1, \ldots, m\}$ such that $C_j \neq D_j$ (and hence $C_k = D_k$ for all $k \neq j$). This edge can and will be identified with the splitting σ_j which determines it. Note that the trivial splitting does not define an edge. As it turns out, all others do. One now has:

9.2.6. Theorem, [6]. Let there be given a set of compatible splittings, Σ, of m elements. Then the definitions (9.2.3) - (9.2.5) define a connected tree of $m+1$

vertices and m edges if $\sigma_0 \notin \Sigma$, *and with m vertices and* $m-1$ *edges if* $\sigma_0 \in \Sigma$.

9.2.7. Remark. Admitting the trivial splitting is a slight departure from Buneman's paper. It makes very little difference, but does make the treatment below a little cleaner.

Now let (M, d) be a metric space (or, more generally, a dissimilarity space). For each splitting $\sigma = \{A, B\}$ of M, define

$$\mu_\sigma = \frac{1}{2} \inf_{a,a' \in A;\ b,b' \in B} \{d(a,b) + d(a',b') - d(a,a') - d(b,b')\}. \tag{9.2.8}$$

A splitting $\sigma = \{A, B\}$ is admissible if and only if $\mu_0 > 0$. Note that the trivial splitting is always admissible.

9.2.9. Lemma, [6]. Two admissible splittings are compatible.

Now define $\Sigma(M)$ to be the set of all admissible splittings, and the Buneman tree, BT_M, of M as the tree defined by $\Sigma(M)$ according to Theorem 9.2.6.

For each vertex $v = \{C_1, \ldots, C_m\}$of BT_M, define its support by

$$\mathrm{supp}(v) = \bigcap_{j=1}^{m} C_j \tag{9.2.10}$$

This support can well be empty. Finally for each $a \in M$ let $\beta(a)$ be the unique vertex in BT_M characterized by

$$\beta(a) = v = \{C_1, \ldots, C_m\} \Longleftrightarrow a \in C_i,\ i = 1, \ldots, m. \tag{9.2.11}$$

This defines a canonical map

$$\beta : M \to BT_m. \tag{9.2.12}$$

Finally let d_T be the network metric on BT_M determined by giving edge σ weight μ_σ, and let d_B be the dissimilarity on M defined by

$$d_B(x,y) = d_T(\beta(x),\ \beta(y)) \tag{9.2.13}$$

Because β need not be injective it can happen that d_B takes the value zero for two unequal points x, y; otherwise it satisfies, of course, the triangle condition and the four-point condition.

9.2.14. Theorem, [6]. *For all* $x, y \in M$, $d_B(x,y) \leq d(x,y)$. *If* d *is a tree metric, i.e. if it satisfies the four-point condition,* $d_B = d$.

This concludes the summary of [6] needed for what will follow.

The whole construction has a very canonical feel to it and has the flavour of

FIGURE 13.

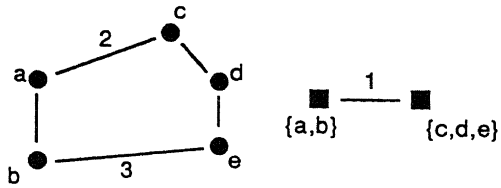


FIGURE 14.

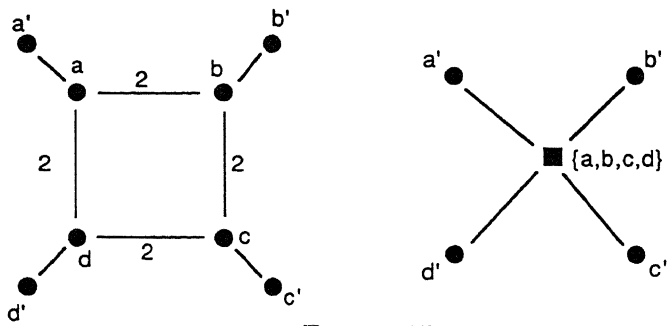


FIGURE 15.

the kind of thing that solves a universal problem of some kind. In contrast to the case of ultrametric spaces, cf 7.5 above, I have no idea of what would be the right categories and morphisms for this in this case.

9.3. Examples
In each of the examples below the original metric space is given first as a network, and its Buneman tree is next to it or below it. In the network representation of the metric space all edges have length 1 unless something else is explicitely indicated.

A vertex of the Buneman tree whose support is precisely one element is depicted by a filled black circle, a vertex with empty support is depicted by an open circle and a vertex whose support consists of more than one element is given by a small black filled square.

– In the Buneman tree on the right of Figure 13, all edges have weight 1/2,

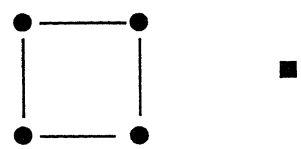


FIGURE 16.

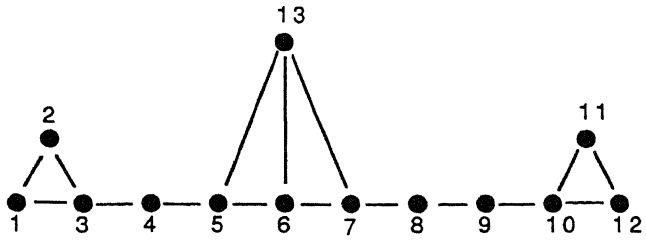


FIGURE 17.1.

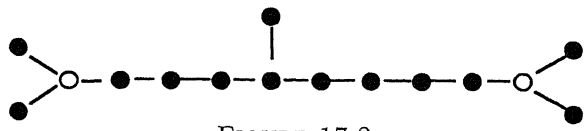


FIGURE 17.2.

so that the original metric is exactly reproduced in this case.

- Figure 14 shows that the canonical map need not be injective. It also shows that the new dissimilarity d_B need not be the suboptimal four-point dissimilarity with respect to d. In this case an edge distance 2 between the two vertices of the Buneman tree would still be uniformly smaller than d. This can also happen if the canonical map is injective.
- In the right half of Figure 15 the four edges all have weight 1.
- For the metric space of Figure 16 there is no other admissible splitting than the trivial one. This is one good reason to admit the trivial splitting.
- Finally in Figure 17 the Buneman tree is given of the standard example, that has been discussed before from various points of view (Figure 5, Figure 12.1). In Figure 17.2 all edges (including the 'vertical' one) have weight 1 except the ones around the empty support vertices on the left and right; the edges from these two groups of three all have weight 1/2.

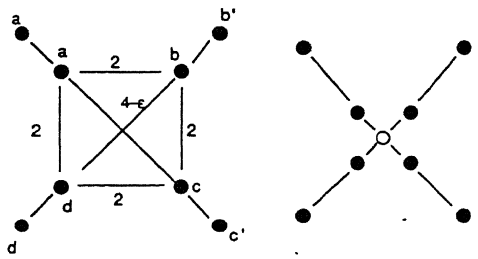


FIGURE 18.

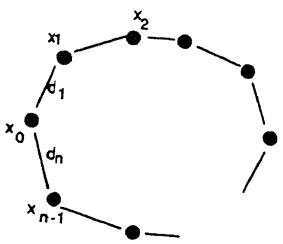


FIGURE 19.

9.4. Continuity
The Buneman tree construction is continuous. More precisely the assignment $d \mapsto d_B$ is continuous. This is easily seen. Indeed every admissible splitting for a given d will remain admissible under a sufficiently small perturbation of d. Also the weight of the corresponding edge will change less than 2ε if the original distance function changes less than ε. It may happen that additional splittings become admissible, as in the example of Figure 18. However, the corresponding edges will have lengths less than 2ε. As an example consider again the network of Figure 15. But now let the diagonal distances, i.e. the distances from a to c and from b to d be $4-\varepsilon$ instead of 4 (and correpondingly the distance from a' to c is $5-\varepsilon$, etc.). Then the Buneman tree changes to the one depicted in Figure 18 where the middle four edges have weight $\frac{1}{2}\varepsilon$.

This example also illustrates that even when the canonical map is injective, d_B need not be a maximal four-point distance that is uniformly smaller than the original distance.

9.5. Locality and circuits
Let us for this occasion define a circuit in a metric space as a sequence of points $x_0, x_1, \ldots, x_{n-1}, x_n = x_0$, with distances $d(x_{i-1}, x_i) = d_i$, such that the distance from x_i to x_j, $i < j$, is either equal to $d_{i+1} + \ldots + d_j$, or equal to $D - (d_{i+1} + \ldots + d_j)$, where $D = d_1 + d_2 + \ldots + d_n$. Now consider a metric space that consists just of one circuit (as in Figure 19). The canonical map from such a circuit space to its Buneman tree will be injective if and only if $2(d_{\min} + d_{\max}) > D$, where and $d_{\min} = \min\{d_1, \ldots, d_n\}$ and $d_{\max} = \max\{d_1, \ldots, d_n\}$. The presence of such circuits seems to be the major reason why the canonical map from a space to its Buneman tree need not be injective. It also seems that the construction of the Buneman tree is more or less localized where 'local' is to be understood in terms of minimal circuits of length 4 or more.

9.6. The strict triangle inequality case
Let M be a metric space that satisfies the strict triangle inequality for all triples of points. Then for all $x \in M$, $\sigma_x = \{\{x\}, M\backslash\{x\}\}$ is an admissible splitting. It follows that the canonical map is always injective in this case. It is also clear that for all $x \in \beta$, $\beta(x)$ will be a leaf of the Buneman tree whose sole incident edge is precisely σ_x. Indeed there can clearly be only one vertex $v = \{C_1, \ldots, C_m\}$ which has at σ_x the component $\{x\}$ (because we must have $C_i \cap C_j \neq \emptyset$ for all i, j). These remarks also solves the following question. Which metrics are such that the four-point condition holds as well as the strict triangle inequality. These are precisely the tree metrics induced on the set of leaves of a tree.

9.7. Perturbations
The easiest way to ensure that the canonical map of a metric space to its Buneman tree is injective is to make sure that the strict triangle inequality holds. So the question arises whether that can be done systematically. This is

in fact very easy. Let ϑ be a function from the nonnegative real numbers to the nonnegative numbers such that

$$\begin{aligned} &\vartheta(0) = 0 \\ &\vartheta(x) > \vartheta(y) \qquad \text{if} \quad x > y \text{ (monotonicity)} \\ &\vartheta(\lambda x) < \lambda\vartheta(x) \quad \text{if} \quad \lambda > 1 \text{ (sublinearity)} \end{aligned} \tag{9.7.1}$$

Then $\vartheta(a+b) < \vartheta(a) + \vartheta(b)$, and it follows that

$$d_\vartheta(x,y) = \vartheta(d(x,y)) \tag{9.7.2}$$

defines a new metric on M for which the strict triangle inequality always holds. Of such functions ϑ there are many. For instance any monotone convex function taking zero to zero such as or $a \mapsto \sqrt{a}$ or $a \mapsto \log(1+a)$. (But a (differentiable) function that satisfies (9.7.1) need not be convex.) It is perhaps even easier to do the following. Let again (M,d) be a metric space. Now change the metric d as follows

$$\begin{aligned} &d_\eta(x,x) = 0 \\ &d_\eta(x,y) = d(x,y) + \eta \text{ if } x \neq y \end{aligned} \tag{9.7.3}$$

where η is a small positive number. Then d_η is a new metric on M (which differs but little from the original one) for which the strict triangle inequality holds. It is possible to describe exactly what happens to the Buneman tree of M under the change $d \mapsto d_\eta$. This needs the following lemma which says that for a metric space it is never just a triangle equality that prevents a splitting from being admissible. (For dissimilarity spaces this lemma need not hold).

9.7.4. Lemma. Let M be a metric space and let $\sigma = \{A, B\}$ be a splitting such that both A and B have two or more elements. Suppose that for all foursomes of different elements $a, a' \in A$, $a \neq a'$; $b, b' \in B$ $b \neq b'$

$$d(a,b) + d(a',b') > d(a,a') + d(b,b') \tag{9.7.5}$$

Then σ is admissible.

PROOF. Because of (9.7.5), if σ were not admissible, this must be the case because there is an $a \in A$ such that

$$d(a,b) + d(a,b') = d(b,b') \tag{9.7.6}$$

for certain $b, b' \in B$(or a similar situation with A and B interchanged, which is treated similarly).
Besides (9.7.5) we also have

$$d(a',b) + d(a,b') > d(a,a') + d(b,b') \tag{9.7.7}$$

Adding this to (9.7.5), and using (9.7.6), we obtain

$$d(a',b) + d(a',b') - d(b,b') - 2d(a,a') > 0 \tag{9.7.8}$$

On the other hand by the triangle inequality

$$d(a',b) \leq d(a,a') + d(a,b)$$

$$d(a',b') \leq d(a,a') + d(a,b')$$

Adding these two, and using (9.7.6) again, gives

$$d(a',b) + d(a',b') \leq 2d(a,a') + d(b,b')$$

which is in conflict with (9.7.8). □

Now let me describe what happens to the Buneman tree of a metric space (M,d) when d is changed to d_η. In any case for all $x \in M$, the splitting $\sigma_x = \{\{x\}, M\backslash\{x\}\}$ becomes admissible if it were not already so. And this is the only thing that happens. Indeed for pairs $a \neq a', b \neq b'$ the admissibility condition does not change when d is replaced by d_η and the lemma says that for splittings that are not of the form σ_x a triangle equality cannot be the only cause of nonadmissibility. It is now not difficult to figure out what happens to the Buneman tree itself. The result is:

- An interior vertex with empty support remains as before (complete with incident edges)
- An interior vertex with nonempty support gets replaced with an interior vertex with empty support. The original incident edges with the original vertex now become incident with the new empty support vertex. In addition there appear as many extra leaves as there were support elements and these are all linked to the new empty support vertex. This is illustrated in Figure 20.

As an example consider the metric space given by the network depicted on the left side of Figure 21. The Buneman tree of the original metric space is in the centre and the Buneman tree of the perturbed metric space is on the right.

9.8. Buneman clusters

It seems to me that the Buneman tree of a metric space gives valuable information as what clusters might be worth consideration. One could for instance find a Buneman tree like the one depicted in Figure 22. This one suggests the various clusters (not all at the lowest level) which are indicated. A precise definition of what clusters are defined by the Buneman tree of a metric space needs to be given.

9.9. Open problems concerning tree metrics

In my opinion the Buneman construction is likely to prove important. Let me list a variety of open questions concerning it.

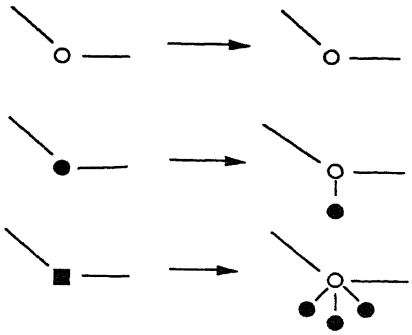


FIGURE 20.

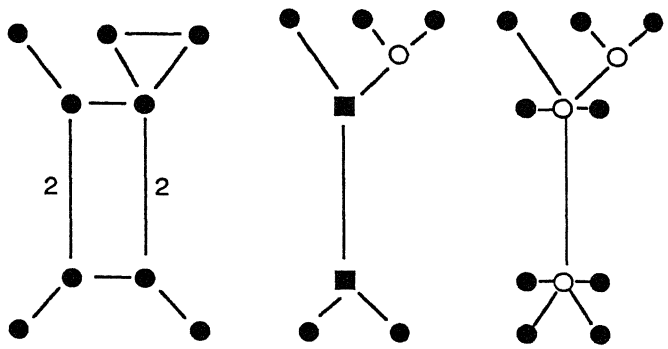


FIGURE 21.

A. The construction has a functorial flavour. It is however far from obvious what should be the categories and classes of morphisms to bring out this functoriality.

B. The Buneman metric d_B of a metric space (M, d) is always smaller than the original metric. It is however not necessarily maximal amoung the tree metrics that are smaller than d. The problem thus arises to find the maximal tree metrics smaller than a given one. In [51], ROUX gives a convergent algorithm for changing a metric into a tree metric which could be examined in this setting. Is there a suboptimal tree metric?

C. Are there systematic ways of constructing the Buneman tree of a network by collapsing parts of minimal circuits and the like?

D. If the metric space satifies the strict triangle inequality (which can always be ensured by a arbitray small perturbation, cf 9.7 above), the original space identifies with the leafs of the Buneman tree. This gives a first level clustering; the clusters are those groups of leafs that have an immediate common ancestor. Now define a metric between these clusters by taking the Hausdorff distance and repeat. This gives a classification scheme. What are its properties?

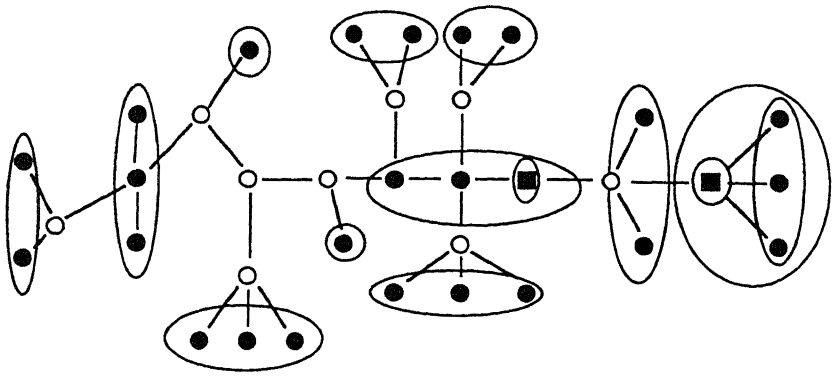


FIGURE 22.

10. OVERLAPPING CLUSTERING

For overlapping clusterings, hierarchical or not, one asks for coverings of the set M instead of partitions. A *covering* of M is a set of subsets $\gamma = \{V_1, V_2, \cdots, V_m\}$ such that

$$\cup_{i=1}^{m} V_i = M \tag{10.0.1}$$

There is a totally trivial way of obtaining a 'classfication scheme' for a dissimilarity space (M, d) which still has its points.

10.1. The poset of balls

For each h define a *ball* of diameter h in M as a maximal subset $B(h)$ such that

$$d(x, y) \leq h, \quad \forall x, y \in B(h) \tag{10.1.1}$$

The maximality here means that if $z \notin B(h)$, then there is an $x \in B(h)$ such that $d(x, z) > h$. The partially ordered set (poset) of balls of (M, d) is now the collection of all balls ordered by inclusion. It comes with a function on it which to each ball associates its diameter. The metric (or dissimilarity) can be easily recovered from these data: the distance between two points is the diameter of the smallest ball in which they are both contained. This is just about totally trivial.

In the example of Figure 23 the distances of the four sides of the diamond are 4, the two vertical pieces are 3 and the two horizontal ones are 2. Thus, the construction is trivial but in concrete cases this may well be the only thing that needs to be done. In any case listing all finest level balls with their members and the inclusion relations between balls can be a vastly more efficient way of describing a metric space (compared to giving the distance function) when the overlap between balls at each level is modest like, say, 10%. Take e.g. a constant depth 4 situation with some 4000 leaves, 20 first generation children,

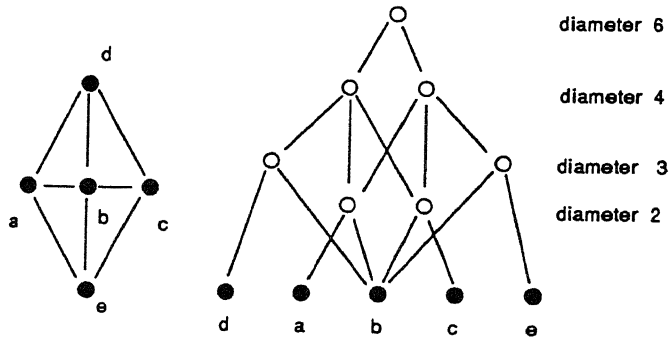


FIGURE 23.

each of which has on average some 10 second generation children; these have in turn on average 6 children and those have again some 6 children. These are the kind of parameters one finds in actual existing classification schemes. Assuming 10% overlap, the ball poset description will be a factor of about 10^4 more efficient. The situation is anlogous to the socalled 'watchmaker paradox'.

Not all partially ordered sets can arise in this way. They obviously have the following property. For each element x of a poset P with largest element, define

$$m(x) = \{y \in P : y \prec x, \quad y \text{ is minimal in } P\} \tag{10.1.2}$$

Then the poset of balls of a dissimilarity space satisfies

$$y \prec x \Leftrightarrow m(y) \subset m(x) \tag{10.1.3}$$

10.2. Powerset metrics

Let M be a set and $Pow(M)$ the set of all subsets of M. Let X be a subset of $Pow(M)$ which includes all one element subsets, and let $t : X \to \mathbb{R}$ be a function such that

$$\begin{aligned} &t(V) \geq 0 \text{ for all } V \in X \\ &t(V) = 0 \text{ if and only if } \#V = 1 \\ &t(V_1) > t(V_2) \text{ if } V_1 \supset V_2 \text{ and } V_1 \neq V_2 \end{aligned} \tag{10.2.1}$$

These data define a metric on M as follows. Construct the network with vertices the elements of X and with edges between V and W if and only if $V \subset W$, $V \neq W$, and there are no elements of X strictly in between V and W. Give this edge the weight $t(W) - t(V)$. Now identify M with the single

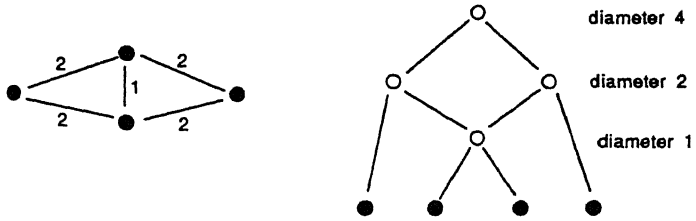


FIGURE 24.

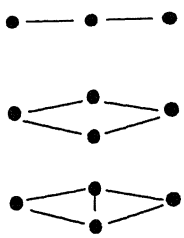


FIGURE 25.

element subsets of $Pow(M)$ (which are all in X), and take the metric on M induced by that network metric of X. I will call such metrics on M *powerset metrics*.

Now take a metric space (M, d) and take for X the poset of balls of it. Define $t : X \to \mathbb{R}$ by $t(B(h)) = \frac{1}{2}h$. This defines a new metric, d_t, on M which need not be the old one as the example of Figure 24 shows. In this example the d_t of the leftmost and rightmost point becomes 3 (while it was 4 originally).

Completely open is the question of which metric spaces are such that the powerset metric defined by their poset of balls is equal to the original one.

Also open is the question of what special properties powerset metrics have. For example the metric space of the left half of Figure 24 cannot come from a powerset metric for which the set X consists of balls.

11. THIRD PARTY SUPPORT

We have seen that in the case of the example that has been considered several times (Figure 5, Figure 12.1) single link clustering is just as good as several others. Still some of the others definitely look better.

In our intended applications the distance function comes from the number of cocitations of two terms or phrases. That is, it comes from a direct comparison of the two items involved while ignoring more indirect relatedness. I would submit that the two outer points in the middle part of Figure 25 are more related than the two outer points in the upper part and that this is still more the case in the lower part. This is what is intended with the phrase 'third party

support'.

Let us think in terms of networks (which is not really a limitation at all).

For two points , define their relatedness as

$$r(x,y) = \sum_p l(p)^{-2} \tag{11.0.1}$$

where the sum is over all paths from x to y without loops. And then define a new distance

$$d_r(x,y) = r(x,y)^{-2} \tag{11.0.2}$$

Now perform single link clustering on the new space (M, d_r). This is one way of trying to incorporate the idea of indirect relatedness. It will be of interest to try this out on real data. Note that the powerset metric d_t defined by the poset of balls also incorporates something of the idea of third party support (cf Figure 24).

What exponent is taken in (11.0.1) and (11.0.2) is largely open. Probably at least two to avoid something like an analogue of the Olber paradox from astronomy.

12. Gravity clustering and other untried ideas

Here are some more possibilities that I feel should be tried out and that I have not found in the literature.

12.1. Gravity clustering

Define the weight of a point $x \in M$ by

$$w(x) = \sum_{x \neq y \in M} d(x,y)^{-2} \tag{12.1.1}$$

A sink is a point whose weight is greater than the weights of all its nearest neighbours. In this proposed scheme these sinks determine clusters. They are the basins of attraction of the sinks. Let the sinks be $x_1, x_2, \ldots, x_n$. Then $y \in M$ belongs to cluster i iff

$$w(x_i)d(y,x_i)^{-2} = \max_{j=1,\ldots,n} \{w(x_j)d(y,x_j)^{-2}\} \tag{12.1.2}$$

(This can cause mild overlap).

To get a hierarchy repeat with as the new data set the clusters formed provided with Hausdorff distance.

Here also, the exponent to be used can be changed.

12.2. Cut-set clustering

A cut-set in a network is a collection of edges such that their removal causes M to decompose into several connected components. Define the strength of a cut-set E as

$$s(E) = \sum_{e \in E} l(e)^{-1} \tag{12.2.1}$$

where $l(e)$ is the length of the edge e. Now proceed as follows. Find the smallest s for which there is a cut-set of strength s. Remove all edges which are in cut-sets of strength s. This defines the first level clusters. Now repeat with each cluster to get the next partitioning. Or go on to the next strength of cut-sets.

13. Dual clustering and bipartite graphs

So far I have exclusively discussed matters from the point of view of clustering objects from one single space, which is thought of as a space of scientific key words and phrases. This does not reflect the background situation completely. In fact we have two spaces: a space of key items and a space of documents and they are related as a bipartite graph with a key item and a document linked if and only if that key item occurs in that document. Moreover the two metrics on these two (dual) spaces are both derived from the incidence matrix, i.e from this bipartite graph.

Thus a good clustering of one of the spaces should tell us things about the other space, and the two clustering problems, one for documents, one for key items, should be treated together. The desirability of doing this is discussed in [11].

Concerning the mathematical problems that arise in such a setting virtually nothing has been done so far. But see [7] for one initial idea.

In particular, there is a well-known other technique for clustering documents, viz co-citation analysis. This leads for instance to research fronts, [50, 55, 60]. These research fronts in turn should give rise to information on key item clustering. This has not yet been explored at all.

References

1. V. Balaz, J. Koca, M. Kvasnicka and M. Sekanina, 1986, Metric for graphs, *Cas. Pest. Mat.* **111**, 431–433.
2. V. Balaz, M. Kvasnicka and J. Popischal, 1989, Dual approach for edge distance between graphs, *Cas. Pest. Mat.* **114**, 155–159.
3. Rodrigo A. Botafogo and Ben Sneiderman, 1991, Identifying aggregates in hypertext structures. In: John J Leggett (ed.), *Proceedings of the third ACM congerence on hypertext*, San Antonio, Dec. 1991, ACM Press, 63–74.
4. P.D. Bruza and L.C. van der Gaag, 1992, *Index expression belief networks for information disclosure*, Preprint, Dept. Comp. Sci., Utrecht Univ.
5. F. Buekenhout (ed.), 1994, *Handbook of incidence geometry*, Elsevier.
6. Peter Buneman, 1971, The recovery of trees from measures of dissimilarity. In: F.R. Hodson, D.G. Kendall and P. Tautu (ed.), *Mathematics in the archeological and historical sciences*, Edinburgh Univ. Press, 387–395.
7. Fuzli Can and Esen A. Ozkarahan, 1982, *A clustering scheme*, ACM, 115–121.

8. GARY CHARTRAND, FARROKH SABA and HUNG-BIN ZOU, 1985, Edge rotations and distance between graphs, *Cas. Pest. Mat.* **110**, 87–91.
9. ZHENMIN CHEN and JOHN W. VAN NESS, 1994, Space-contracting, space dilating, and positive admissible clustering algorithms, *Pattern Recognition* **27**:6, 853–857.
10. DG XIII/E COMMISSION OF THE EUROPEAN COMMUNITIES, 1993, *Strategic study on New opportunities in the information services market*, European Union.
11. DONALD B. CROUCH, 1975, A file organization and maintenance procedure for dynamic document collections, *Information Processing and Management* **11**, 11–21.
12. WILLIAM H.E. DAY and HERBERT EDELSBRUNNER, 1984, Efficient algorithm for agglomerative hierarchical clustering methods, J. Classification 1:7, 7–24.
13. ANITA DE WAARD, 1993, FOM-thema dag belicht remedies informatica infarct, *Ned. Tijdschrift voor Natuurkunde*: **8**, 127–128.
14. EDWIN DIDAY, CHIKIO HAYASHI, MICHEL JAMBU and NOBORU OHSUMI (ed.), 1988, *Recent developments in clustering and data analysis*, Academic Press.
15. ANNETTE J. DOBSON, 1974, Unrooted trees for numerical taxonomy, *J. applied Prob.* **11**, 32–42.
16. LLOYD FISHER and JOHN W. VAN NESS, 1971, Admissable clustering procedures, *Biometrica* **58**:1, 91–104.
17. KENJI FUKAYA, 1986, On a compactification of the set of Riemannian manifolds with bounded curvatures and diameters. In: *Curvature and topology of Riemannian manifolds.* Proceedings of the 17-th international Taniguchi symposium, Springer, 89–107.
18. ERHARD GODEHARDT, 1990, *Graphs as structural models*, Vieweg.
19. J.G. GREGORY, 1983, Citation study of peripheral theories in an expanding research front, *J. Information Science* **7**, 71–80. Present
20. M. GROMOV, 1981, *Structures métriques pour les variétés Riemanniennnes*, Cedic/Fernand Nathan.
21. A. GUÉNOCHE, P. HANSEN and B. JAUMARD, 1991, Efficient algorithms for divisive hierarchical clustering with the diameter criterion, *J. Classification* **8**, 5–30.
22. PIERRE HANSEN and MICHEL DELATTRE, 1978, Complete-link cluster analysis by graph coloring, *J. American Stat. Ass.* **73**, 397–403.
23. J.A HARTIGAN, 1967, Representation of similarity matrices by trees, *J. American Stat. Ass.* **62**, 1140–1158.
24. M. HAZEWINKEL (ed.), 1988-1994, *Encyclopaedia of mathematics*, 10 volumes, KAP.
25. MICHIEL HAZEWINKEL, 1991, *The chart and web of mathematics*, Research proposal, CWI.
26. MICHIEL HAZEWINKEL, 1992/1993, *Mathematical information, Notes of lectures.*
27. MICHIEL HAZEWINKEL, 1993, *Notes on the 'Encyclopaedia of Mathematics'*

component of the 'Chart and Web of Mathematics'.
28. MICHIEL HAZEWINKEL, 1994, *BUC'M, bottom-up classification in mathematics and physics*, Research proposal.
29. MICHIEL HAZEWINKEL, 1994, *Infinite scientific information*, Notes and slide copies of a lecture at KAP.
30. MICHIEL HAZEWINKEL and CLYDE F. MARTIN, 1983, Representations of the symmetric groups, the specialization order, Schubert cells, and systems, *Enseignement Math.* **29**, 53–87.
31. ERIC W. HOLMAN, 1992, Statistical properties of large published classifications, *J. Classification* **9**, 187–210.
32. LAWRENCE J. HUBERT, 1974, Some applications of graph theory for clustering, *Psychometrika* **39**:3, 283–309.
33. M. JAMBU and M.-O. LEBEAUX, 1983, *Cluster analysis and data analysis*, North Holland.
34. NICHOLAS JARDINE and ROBIN SIBSON, 1971, *Mathematical taxonomy*, Wiley.
35. S.C. JOHNSON, 1967, Hierarchical clustering schemes, *Psychometrika* **32**, 241–254.
36. FRIEDER KADEN, 1990, Distance graphs on finite graph sets, *Rostock Math. Kolloq.* **41**, 39–46.
37. M. KATETOV, 1988, On universal metric spaces. In: Z. FROLIK (ed.), General topology and its relations to modern algebra and analysis VI, *Heldermann Verlag*, 323–330.
38. RAY R. LARSON, 1991, Classification clustering: probabilistic information retrieval, and the online catalogue, *Library Quarterly* **61**:2, 133-173.
39. D. LEUSCHNER, 1991, A mathematical model for classification and identification, *J. Classification* **8**, 99-113.
40. JIRI MATOUSEK, 1990, Bi-Lipschitz embeddings into low dimensional Euclidean spaces, *Comm. Math. Univ. Carolinae* **31**:3, 589-600.
41. HANS-JOACHIM MUCHA, 1992, Clusteranalyse mit Mikrocomputern, *Akademie Verlag.*
42. F. MURTAGH, 1983, A probability theory of hierarchic clustering using random dendograms, *J. Stat. Comp. Simul.* **18**, 145–157.
43. F. MURTAGH, 1983, A survey of recent advances in hierarchical clustering algorithms, *Computer Journal* **26**:4, 354–359.
44. F. MURTAGH, 1984, Counting dendograms: a survey, *Discr. Appl. Math.* **7**, 191-199.
45. F. MURTAGH and A. HECK, 1987, *Multivariate data analysis*, Reidel.
46. I.G. NIKOLAEV, 1989, On the closure of the set of classical Riemannian spaces, Itogi Nauki Tekh., *Ser. Probl. Geom.* **21**, 43–66.
47. J.S. OTTAVIANI, 1994, The fractal nature of relevance, *J. Amer. Soc. Information Sci.* **45**:4, 263–272.
48. JAN W. OWSINSKI, 1990, On a new naturally indexed quick clustering method with global objective function, *Applied Stochastic Models and Data Analysis* **6**, 157-171.
49. A.N. PATRINOS and S.L. HAKIMI, 1972, The distance matrix of a graph

and its tree realization, *Quarterly of applied Math.* **30**, 255–269.
50. OLLE PERSSON, 1994, The intellectual base and research fronts of JASIS 1986–1990, *J. Amer. Soc. Inf. Sci.* **45**:1, 31-38.
51. M. ROUX, 1988, *Techniques of approximation for building two tree structures.* In: EDWIN DIDAY, CHIKIO HAYASHI, MICHLE JAMBU and NOBORU OHSUMI (ed.), Recent developments in clustering and data analysis, Academic Press, 151-170.
52. MICHAL SABO, 1991, On a maximal distance between graphs, *Chechoslovak Math. J.* **41**, 265–268.
53. PETER H.A. SNEATH and ROBERT R. SOKAL, 1973, *Numerical taxonomy*, Freeman.
54. FRED SOBIK, 1982, On some measures of distance between graphs. In: Graphs and other combinatorial topics. *Proceedings of the third Chechoslovak symposium*, Teubner.
55. MASAYA TAKAYAMA, 1986, Analysis of technological information transfer among Japanese computer scientists at a research front, *Information Services and Use* **6**, 9-25.
56. P. URYSOHN, 1925, Sur un espace métrique universel, *Comptes Rendus Acad. Sci. Paris* **180**, 803–806.
57. P. URYSOHN, 1927, Sur un espace métrique universel, *Bull. Sci. Math.* **51**, 43–64.
58. V.V. USPENSKIJ, 1990, On the group of isometries of the Urysohn universal metric space, *Comm. Math. Univ. Carolinae* **31**:1, 181–182.
59. JOHN W. VAN NESS, 1973, Admissable clustering procedures, *Biometrica* **60**, 422–424.
60. A.F.J. VAN RAAN and R.J.W. TIJSSEN, 1992, The neural net of neural research, *Scientometrics* **26**:1, 169–192.
61. B. ZELINKA, 1987, Distance between isomorphism classes of graphs, *Cas. Pest Mat.* **112**, 233–237.