Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Lipshits distance and hierarchical clustering

M. Hazewinkel

Department of Analysis, Algebra and Geometry

AM-R9506 1995

Report AM-R9506
ISSN 0924-2953

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Lipshits distance and hierarchical clustering

Michiel Hazewinkel
CWI
POBox 94079
1090GB Amsterdam
The Netherlands
mich@cwi.nl

Abstract

This note discusses the Lipshits distance between two different metrics (or definite) dissimilarities on a set. Given a metric space a universal lower bound is established for the Lipshits distance between the original metric on M and the metric defined by any hierarchical classification tree on M. Finally it is shown that single link clustering attains this lower bound.

Mathematics Subject Classification 1991: 62H30, 68T10, 54E35.

Key words & Phrases: Lipshits distance, hierarchical clustering, single link clustering.

1. Classification trees. A definite dissimilarity space is a (finite) set M together with a function $d : M \times M \to \mathbf{R}$ such that $d(x,y) = d(y,x) > 0,\ d(x,x) = 0$ for all $x, y \in M, x \neq y$. This function is called a dissimilarity. If in addition the triangle inequality is satisfied, (M,d) is a metric space.

In this note a hierarchical classification tree for M is a series of coarser and coarser partitions

$$\{\text{singletons}\} = \pi_0 \ \mathrm{p}\ \pi_1 \ \mathrm{p}\ \pi_2 \ \mathrm{p}\ \mathrm{L}\ \ \mathrm{p}\ \pi_m = \{M\} \tag{1.1}$$

together with a (level) function on the set of partitions in the form of m strictly positive numbers

$$d_1 < d_2 < \mathrm{L}\ \ < d_m . \tag{1.2}$$

These data define a new metric on M by

$$d_T(x,y) = d_j \tag{1.3}$$

if and only if j is the smallest integer such that both x,y are in a same set of the partition π_j.

Inversely, given an ultrametric which assumes the values $d_i, i = 1, \mathrm{L}\ , m$, the balls with diameter d_i (maximal sets all of whoce members have distance less than d_i to one another) define a partition π_i; the corepondence thus set up is bijective [Hartigan, 1967; Johnson, 1967]. In this note the sequence of partitions point of view is more convenient for the proofs and constructions.

2. Lipshitz distance.
Consider a set M and two metrics (or dissimilarities), d_1, d_2 defined on it. The *distortion* of d_2 with respect to d_1 is defined by

$$\mathrm{distor}(d_2, d_1) = \max \frac{d_2(x,y)}{d_1(x,y)} \tag{2.1}$$

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

$$\delta_L(d_1, d_2) = \log(\text{distor}(d_2, d_1)\text{distor}(d_1, d_2)). \tag{2.2}$$

Note that if the two distances are proportional, their Lipshits distance is zero. This feature is really an advantage for classification problems because a constant scalar factor should not matter.

It is easy to see that:

4.2. *Proposition* [Johnson, Lindenstrauss, and Schechtman, 1987; Matousek, 1990],. The Lipshitz distance δ_L defines a metric on isometry classes up-to a scalar factor of metrics (or definite dissimilarities) on a fixed set *M*.

3. 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 (sl) of P is equal to

$$sl(P) = \max_{i=0,\ldots,n} \{d(x_i, x_{i+1})\}. \tag{3.1}$$

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

$$sas(x, y) = \min_P \{sl(P)\} \tag{3.2}$$

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

$$dsq(M) = \max_{x \neq y} \frac{d(x, y)}{sas(x, y)}. \tag{3.3}$$

3.4. *Theorem*. Let $\pi = (\pi_0, \pi_1, \ldots \pi_n)$ be a 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 (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

$$\text{distor}(d, d_T) \geq \frac{d(x, y)}{d_j}. \tag{3.5}$$

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 π_{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 $y \notin C_{11}$. 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 greater than *j*, so that

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

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

(3.6) $$\text{distor}(d_T, d) \geq \frac{d_j}{sas(x,y)}.$$ (3.6)

Inequalities (3.5) and 3.6) together prove the theorem.

4. Single link clustering.
Single link clustering yields the following hierarchy. Let $d_1 < d_2 < \mathrm{L}\ d_s$ be the distances that actually occur in the original metric. 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.

4.1. *Theorem*. Single link clustering is optimal with respect to Lipshits distance. More precisely:

(4.2) $$\delta_L(d_{sl}, d) = \log(dsq(M)).$$

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

(4.3) $$\delta_L(d_{sl}, d) \leq \log(dsq(M)).$$

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

(4.4) $$\text{distor}(d, d_{sl}) = \max \frac{d(x,y)}{d_{sl}(x,y)} \leq dsq(M)$$

On the other hand, as is rather obvious and also well known, [Jardine and Sibson, 1971, p. 63], $d_{sl}(x,y) \leq d(x,y)$ for all $x, y \in M$. Hence

$$\text{distor}(d_{sl}, d) \leq 1.$$ (4.5)

The combination of (4.4) and 4.5) establishes (4.3) and hence (4.2) and the theorem.

4.6. *Remark.* In [Jardine, Jardine, and Sibson, 1967] it is shown that d_u is subdominant; i.e. that is it is maximal with respect to the ordering $d \mathrel{\mathrm{p}} d' \Leftrightarrow \forall_{x,y} d(x,y) \leq d'(x,y)$ amoung the ultrametrics that are no greater in this ordering than the original metric. This, however, carries no implications with respect to its Lipshits distance to the original metric.

4.7. *Corollary.* As noted during the proof

$$sas(x,y) = d_{sl}(x,y)$$

5. Example.

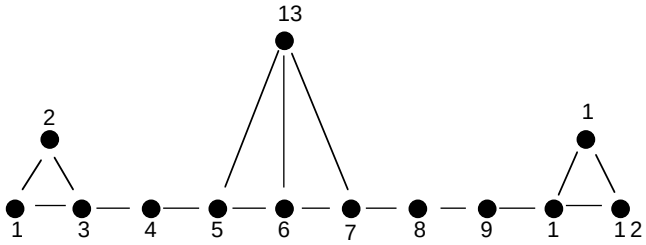

Figure 1

Though the single link classification hierarchy is optimal with respect to Lipshits distance, it need not be the only Lipshits nearest hierarchy. The following example shows this and also illustrates why some intuitively appealing clusterings still have no smaller Lipshits distance to the original metric. Let M be the metric space defined by the network shown above in Figure 1. Here the short edges are of length 1 and the longer ones, i.e. the ones to point 13, have length 2. The distance between two points is the length of the shortest path connecting them. For this space $dsq(M) = 9$. Figure 2 gives the single link hierarchy clustering for this space. Two other clusterings are in Figures 3 and 4. All three of these clusterings have Lipshitz distance log(9) to the original metric. The reason that this is the case in the case of the clusterings of Figures 4 and 5 is that, though the clusters themselves are perhaps better, one pays a high price for breaking up nearest pairs like 4, 5 and 8, 9 in Figure 3 and 4, 5 and 7, 8 in Figure 4.

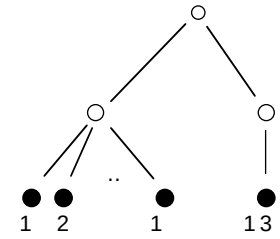

Figure 2

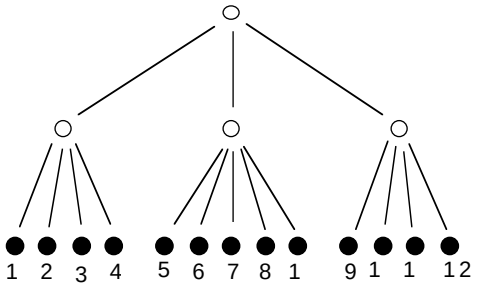

Figure 3

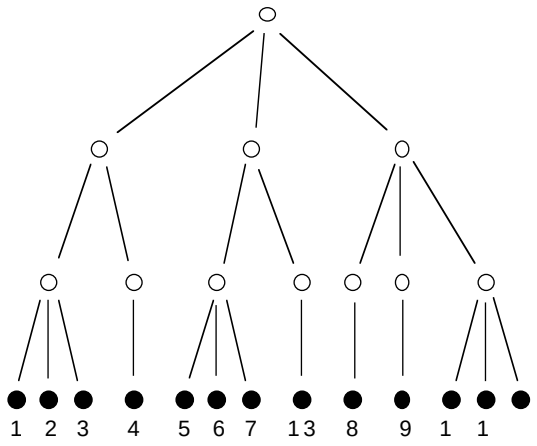

Figure 4

References.

[Hartigan, 1967 #57; Jardine, 1967 #123; Johnson, 1967 #73; Johnson, 1987 #114; Matousek, 1990 #13; Jardine, 1971 #70]

1. N. Jardine, R.Sibson, *Mathematical taxonomy*, Wiley, 1971.

2. W.B. Johnson, J. Lindenstrauss, G.Schechtman, *On Lipschits embeddings of finite metric spaces in low dimensional normed spaces*, In: J Lindenstrauss, V.D. Millman (eds), Geometrical aspects of functional analysis, Springer, 1987.

3. J. Matousek, *Bi-Lipschitz embeddings into low dimensional Euclidean spaces*, Comm. Math. Univ. Carolinae 31:3(1990), 589-600.