A Note on “Performance Guarantees for Hierarchical Clustering”

Jef Wijsen

March 12, 2008

Abstract

This note may help you in understanding Sections 3.3 and 3.4 of [1].

Levels of Granularity

I will use lowercase letters for points, and uppercase letters for levels.

- Point 1 is at level 0.
- For $i \geq 2$, point i is at level J if
  1. $R_i > \frac{R}{\beta^J}$, and
  2. point i is not at level $J-1$ (more precisely, $R_i \not> \frac{R}{\beta^{J-1}}$).

The function $lev(\cdot)$ maps every point to its level. We write L_K for the set of points at level K. Consequently, $i \in L_K$ if and only if $lev(i) = K$. Since i is at level $lev(i)$, it follows

$$R_i > \frac{R}{\beta^{lev(i)}} \tag{1}$$

This is visualized in Fig. 1.

Property 1 $L_0 \cup L_1 \cup \ldots \cup L_J = \{i \mid R_i > \frac{R}{\beta^J}\}$.
Or equivalently, $k \notin L_0 \cup L_1 \cup \ldots \cup L_J$ *implies* $R_k \leq \frac{R}{\beta^J}$.

Assume that $L_0 \cup L_1 \cup \ldots \cup L_J = \{1, 2, \ldots, l\}$. The most distant point from $\{1, 2, \ldots, l\}$ is point $l+1$ at distance R_{l+1}. By Property 1, $R_{l+1} \leq \frac{R}{\beta^J}$. It follows that the distance between any point and (the closest point in) $L_0 \cup L_1 \cup \ldots \cup L_J$ is at most $\frac{R}{\beta^J}$ (this is Lemma 7).

As an immediate consequence, the distance between point i and (the closest point in) $L_0 \cup L_1 \cup \ldots \cup L_{lev(i)-1}$ is at most $\frac{R}{\beta^{lev(i)-1}}$. Since the point of $L_0 \cup L_1 \cup \ldots \cup L_{lev(i)-1}$ that is most close to i is denoted $\pi'(i)$, we have

$$distance(i, \pi'(i)) \quad \leq \quad \frac{R}{\beta^{lev(i)-1}} \tag{2}$$

This is Corollary 8.

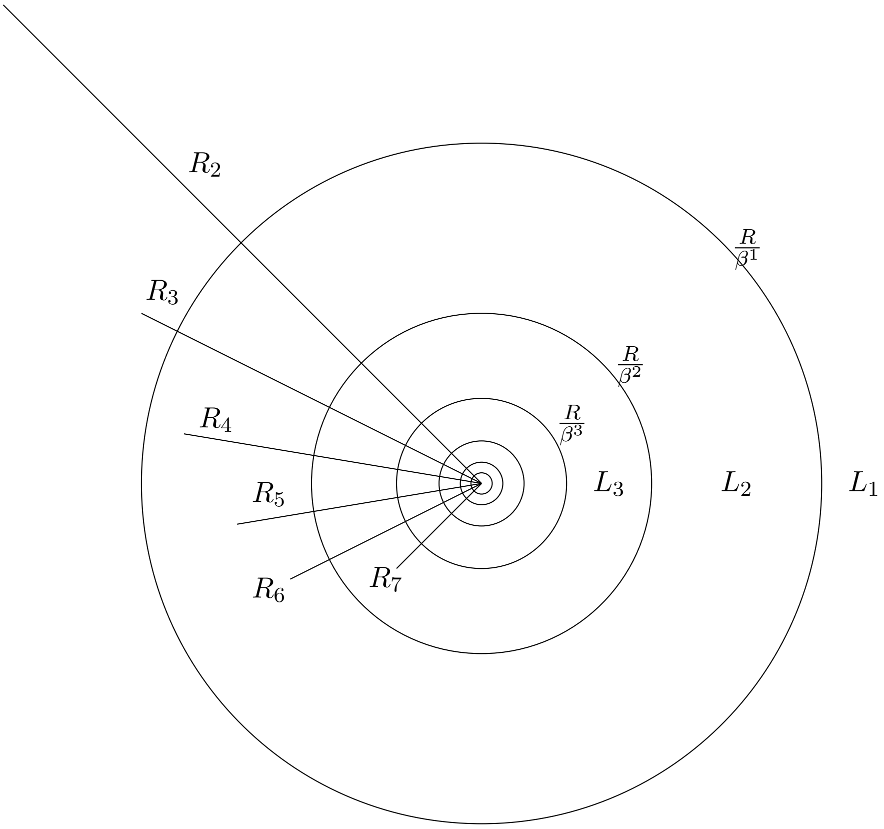

Figure 1: Example distribution of R_i's: $L_1 = \{2, 3\}$, $L_2 = \{4, 5, 6\}$, $L_3 = \{7\}$. The picture uses $\beta = 2$.

A Performance Guarantee

Recall:

- To obtain a 2-clustering, we remove the edge between point 2 and point $\pi'(2) = 1$. The points 1 and 2, which are no longer connected, are taken as cluster centers.
- Then, to obtain a 3-clustering, we furthermore remove the edge between point 3 and point $\pi'(3) \in \{1, 2\}$. The points 1, 2, and 3, which are now mutually disconnected, are taken as cluster centers.
- ...
- Then, to obtain a k-clustering, we furthermore remove the edge between point k and point $\pi'(k) \in \{1, 2, \ldots, k-1\}$. The points $1, 2, \ldots, k$, which are now mutually disconnected, are taken as cluster centers.

For a point $i > k$, to find the center of its cluster, we follow a path

$$i = i_0 > i_1 > i_2 > \ldots > i_{l-2} > i_{l-1} > i_l \in \{1, 2, \ldots, k\} \ ,$$

where $\pi'(i_0) = i_1$, $\pi'(i_1) = i_2$, ..., $\pi'(i_{l-2}) = i_{l-1}$, $\pi'(i_{l-1}) = i_l$.

By (2),

$$\begin{aligned} \mathit{distance}(i_0, i_1) &\leq \frac{R}{\beta^{lev(i_0)-1}} \\ \mathit{distance}(i_1, i_2) &\leq \frac{R}{\beta^{lev(i_1)-1}} \\ &\vdots \\ \mathit{distance}(i_{l-3}, i_{l-2}) &\leq \frac{R}{\beta^{lev(i_{l-3})-1}} \\ \mathit{distance}(i_{l-2}, i_{l-1}) &\leq \frac{R}{\beta^{lev(i_{l-2})-1}} \\ \mathit{distance}(i_{l-1}, i_l) &\leq \frac{R}{\beta^{lev(i_{l-1})-1}} \end{aligned}$$

Since $lev(i_{l-2}) > lev(i_{l-1})$, we have $lev(i_{l-2}) - 1 \geq lev(i_{l-1})$. Consequently,

$$\mathit{distance}(i_{l-2}, i_{l-1}) \leq \frac{R}{\beta^{lev(i_{l-2})-1}} \leq \frac{R}{\beta^{lev(i_{l-1})}} = \frac{1}{\beta} \cdot \frac{R}{\beta^{lev(i_{l-1})-1}} \ .$$

Likewise, $lev(i_{l-3}) - 1 \geq lev(i_{l-2})$. Consequently,

$$\begin{aligned} \mathit{distance}(i_{l-3}, i_{l-2}) \leq \frac{R}{\beta^{lev(i_{l-3})-1}} \leq \frac{R}{\beta^{lev(i_{l-2})}} &= \frac{1}{\beta} \cdot \frac{R}{\beta^{lev(i_{l-2})-1}} \\ &\leq \frac{1}{\beta^2} \cdot \frac{R}{\beta^{lev(i_{l-1})-1}} \ . \end{aligned}$$

By repeated application of the same reasoning, it follows that the total path length is bounded by:

$$(\ldots + \frac{1}{\beta^2} + \frac{1}{\beta} + 1) \cdot \frac{R}{\beta^{lev(i_{l-1})-1}} \leq \frac{\beta}{\beta - 1} \cdot \frac{R}{\beta^{lev(i_{l-1})-1}} \ .$$

Recall that the cost of every k-clustering is at least $\frac{R_{k+1}}{2}$. So it suffices to prove that the total path length is at most $\mathcal{O}(1) \cdot \frac{R_{k+1}}{2}$.

Since $i_{l-1} > k$, we have $i_{l-1} \geq k+1$, hence $lev(i_{l-1}) \geq lev(k+1)$. Consequently,

$$\frac{R}{\beta^{lev(k+1)}} \geq \frac{R}{\beta^{lev(i_{l-1})}} \quad .$$

Since $R_{k+1} > \frac{R}{\beta^{lev(k+1)}}$ by (1), we have $R_{k+1} \geq \frac{R}{\beta^{lev(i_{l-1})}}$. Consequently,

$$\frac{\beta^2}{\beta-1} \cdot R_{k+1} \geq \frac{\beta^2}{\beta-1} \cdot \frac{R}{\beta^{lev(i_{l-1})}} = \frac{\beta}{\beta-1} \cdot \frac{R}{\beta^{lev(i_{l-1})-1}} \quad .$$

It follows that the path length from any point i to its cluster center is at most $\frac{2\beta^2}{\beta-1} \cdot \frac{R_{k+1}}{2}$. If we choose $\beta = 2$, the cost of the k-clustering obtained is at most 8 times the minimal cost.

References

[1] S. Dasgupta and P. M. Long. Performance guarantees for hierarchical clustering. *J. Comput. Syst. Sci.*, 70(4):555–569, 2005.