A Note on Hierarchical Clustering

Jef Wijsen

March 17, 2009

1 Preliminaries

We assume a universe $\mathbb{O}$ of *objects*, equipped with a distance metric dist. That is, $\mathsf{dist} : \mathbb{O} \times \mathbb{O} \rightarrow \mathbb{R}$ such that for all $o, p, q \in \mathbb{O}$:

1. $\mathsf{dist}(o, p) \geq 0$
2. $\mathsf{dist}(o, p) = 0$ iff $o = p$
3. $\mathsf{dist}(o, p) = \mathsf{dist}(p, o)$
4. *Triangle inequality:* $\mathsf{dist}(o, p) \leq \mathsf{dist}(o, q) + \mathsf{dist}(q, p)$

Let $o \in \mathbb{O}$ and $r \in \mathbb{R}$, $r \geq 0$. The *radius r ball around o*, denoted $B_r(o)$, is defined by:

$$B_r(o) = \{p \in \mathbb{O} \mid \mathsf{dist}(o, p) \leq r\}$$

Let $O \subseteq \mathbb{O}$. We define:

$$\begin{aligned} \mathsf{diameter}(O) &= \max\{\mathsf{dist}(o, p) \mid o, p \in O\} \\ \mathsf{radius}(O) &= \min\{r \in \mathbb{R}^+ \mid \exists c \in \mathbb{O} : O \subseteq B_r(c)\} \end{aligned}$$

Theorem 1 *Let $O \subseteq \mathbb{O}$, $|O| \geq 2$. Then,*

$$1 \leq \frac{\mathsf{diameter}(O)}{\mathsf{radius}(O)} \leq 2$$

Proof. Let $d = \mathsf{diameter}(O)$ and $r = \mathsf{radius}(O)$. We can assume $o, p \in O$ such that $d = \mathsf{dist}(o, p)$. Furthermore, we can assume $c \in \mathbb{O}$ such that $O \subseteq B_r(c)$. Since $O \subseteq B_d(o)$, it follows $r \leq d$. From $\mathsf{dist}(c, p) \leq r$, $\mathsf{dist}(c, q) \leq r$, and $d \leq \mathsf{dist}(c, p) + \mathsf{dist}(c, q)$, it follows $d \leq 2r$. □

2 Partitional Clustering

All definitions that follow are relative to some $O \subseteq \mathbb{O}$ with $N = |O|$ and distance metric dist.

Definition 1 Let k be a positive integer such that $1 \leq k \leq N$. A *k-clustering of O* (or simply *clustering* if k and O are understood) is a partition $\{C_1, \ldots, C_k\}$ of O, where

1. for each $i \in \{1, \ldots, k\}$, $\{\} \neq C_i \subseteq O$;
2. for each $i, j \in \{1, \ldots, k\}$ such that $i \neq j$, $C_i \cap C_j = \{\}$; and
3. $\bigcup_{i=1}^{k} C_i = O$.

We use $\mathbb{C}_k$ to denote a k-clustering of O. Every element C_i of a k-clustering $\mathbb{C}_k$ is called a *cluster*. Obviously, $\mathbb{C}_1 = \{O\}$ and $\mathbb{C}_N = \{\{o\} \mid o \in O\}$. For a k-clustering $\mathbb{C}_k$, we define:

$$\mathsf{cost}(\mathbb{C}_k) = \max\{\mathsf{diameter}(C) \mid C \in \mathbb{C}_k\}, \text{ the cost of } \mathbb{C}_k.$$

□

Alternatively, the cost of a clustering could be taken to be the maximal radius of its clusters.

Definition 2 Let

$$\mathsf{optcost}(k) = \min\{\mathsf{cost}(\mathbb{C}_k) \mid \mathbb{C}_k \text{ is a } k\text{-clustering of } O\}$$

A k-clustering $\mathbb{C}_k$ is called *optimal* if $\mathsf{cost}(\mathbb{C}_k) = \mathsf{optcost}(k)$. □

Example 1 Let $O = \{1, 2, 3, 4, 5, 6\} \subseteq \mathbb{N}$. Let $\mathsf{dist}(i, j) = |i - j|$. Then,

- $\mathsf{optcost}(2) = 2$, which is the cost of the 2-clustering $\{\{1, 2, 3\}, \{4, 5, 6\}\}$; and
- $\mathsf{optcost}(3) = 1$, which is the cost of the 3-clustering $\{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$.

□

3 Hierarchical Clustering

Definition 3 Let $\mathbb{C}_k$ and $\mathbb{C}_l$ be two clusterings of the same set with $k > l$. We write $\mathbb{C}_k \prec \mathbb{C}_l$ if for every $C \in \mathbb{C}_k$, there exists $D \in \mathbb{C}_l$ such that $C \subseteq D$. □

Example 2 Let $O = \{1, 2, 3, 4, 5, 6\}$. Let $\mathbb{C}_3 = \{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$. Let $\mathbb{C}_2 = \{\{1, 2, 3\}, \{4, 5, 6\}\}$. Let $\mathbb{C}'_2 = \{\{1, 2, 3, 4\}, \{5, 6\}\}$. Then, $\mathbb{C}_3 \not\prec \mathbb{C}_2$ and $\mathbb{C}_3 \prec \mathbb{C}'_2$. □

Lemma 1 *Let $\mathbb{C}_k \prec \mathbb{C}_l$ with $k > l$ be two clusterings of the same set. For all $C \in \mathbb{C}_k$, $D \in \mathbb{C}_l$,*

$$C \cap D \neq \{\} \iff C \subseteq D \ .$$

Proof. The implication $\Leftarrow$ is trivial. For the opposite implication, assume $C \cap D \neq \{\}$. We can assume $a \in C \cap D$. Since $\mathbb{C}_k \prec \mathbb{C}_l$, we can assume $D' \in \mathbb{C}_l$ such that $C \subseteq D'$. Since $a \in D \cap D'$, we have $D = D'$. Consequently, $C \subseteq D$. □

Lemma 2 *Let $\mathbb{C}_k \prec \mathbb{C}_l$ with $k > l$ be two clusterings of the same set. For every $D \in \mathbb{C}_l$, $D = \bigcup\{C \in \mathbb{C}_k \mid C \subseteq D\}$.*

Proof. Assume $a \in D$. We can assume $C_a \in \mathbb{C}_k$ such that $a \in C_a$. Since $C_a \cap D \neq \{\}$, $C_a \subseteq D$ by Lemma 1. Consequently, $a \in \bigcup\{C \in \mathbb{C}_k \mid C \subseteq D\}$. Since a is an arbitrary element of D, $D \subseteq \bigcup\{C \in \mathbb{C}_k \mid C \subseteq D\}$. The opposite inclusion is trivial. □

Corollary 1 *If $\mathbb{C}_{k+1} \prec \mathbb{C}_k$ are two clusterings of the same set, then for some $C_1, C_2 \in \mathbb{C}_{k+1}$ such that $C_1 \neq C_2$,*

$$\mathbb{C}_k = (\mathbb{C}_{k+1} \setminus \{C_1, C_2\}) \cup \{C_1 \cup C_2\} \ .$$

Proof. Assume $\mathbb{C}_{k+1} \prec \mathbb{C}_k$. Then,

- for every $C \in \mathbb{C}_{k+1}$, there exists a unique $D \in \mathbb{C}_k$ such that $C \subseteq D$; and
- for every $D \in \mathbb{C}_k$, there exists $C \in \mathbb{C}_{k+1}$ such that $C \subseteq D$.

Since $|\mathbb{C}_{k+1}| = |\mathbb{C}_k| + 1$, we can assume w.l.o.g. the following numbering of clusters:

- $\mathbb{C}_{k+1} = \{C_1, \ldots, C_{k+1}\}$ and $\mathbb{C}_k = \{D_1, \ldots, D_k\}$;
- $C_1 \subseteq D_1$, $C_2 \subseteq D_2$, …, $C_k \subseteq D_k$; and
- $C_{k+1} \subseteq D_k$.

By Lemma 2, $D_1 = C_1$, $D_2 = C_2$, …, $D_{k-1} = C_{k-1}$, and $D_k = C_k \cup C_{k+1}$. □

Definition 4 A *hierarchical clustering* (of O) is a sequence

$$\mathbb{H} = \mathbb{C}_N \prec \mathbb{C}_{N-1} \prec \ldots \prec \mathbb{C}_1 \ ,$$

where each $\mathbb{C}_k$ is a k-clustering (of O). □

Notice that a hierarchical clustering is *not* a clustering, but a sequence of clusterings.

Example 3 Two hierarchical clusterings of $\{1,2,3,4,5,6\}$ are as follows:

- $$\begin{aligned} \mathbb{H}_1 = \ & \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\} \\ & \prec \{\{1,2\},\{3\},\{4\},\{5\},\{6\}\} \\ & \prec \{\{1,2\},\{3,4\},\{5\},\{6\}\} \\ & \prec \{\{1,2\},\{3,4\},\{5,6\}\} \\ & \prec \{\{1,2,3,4\},\{5,6\}\} \\ & \prec \{\{1,2,3,4,5,6\}\} \end{aligned}$$
- $$\begin{aligned} \mathbb{H}_2 = \ & \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\} \\ & \prec \{\{1\},\{2\},\{3\},\{4,5\},\{6\}\} \\ & \prec \{\{1\},\{2,3\},\{4,5\},\{6\}\} \\ & \prec \{\{1\},\{2,3\},\{4,5,6\}\} \\ & \prec \{\{1,2,3\},\{4,5,6\}\} \\ & \prec \{\{1,2,3,4,5,6\}\} \end{aligned}$$

Notice that $\mathbb{H}_1$ contains the unique optimal 3-clustering (call it $\mathbb{C}_3^o$), and $\mathbb{H}_2$ contains the unique optimal 2-clustering (call it $\mathbb{C}_2^o$). Since $\mathbb{C}_3^o \not\prec \mathbb{C}_2^o$, there exists no hierarchical clustering that contains both $\mathbb{C}_3^o$ and $\mathbb{C}_2^o$. □

There are two main approaches to construct a hierarchical clustering $\mathbb{C}_N \prec \mathbb{C}_{N-1} \prec \ldots \prec \mathbb{C}_1$:

Agglomerative: each $\mathbb{C}_k$ is constructed from $\mathbb{C}_{k+1}$, starting from $\mathbb{C}_N$.

Divisive: each $\mathbb{C}_k$ is constructed from $\mathbb{C}_{k-1}$, starting from $\mathbb{C}_1$.

4 Problem Statement

Dasgupta and Long [DL05] give a divisive algorithm for constructing a hierachical clustering $\mathbb{C}_N \prec \mathbb{C}_{N-1} \prec \ldots \prec \mathbb{C}_1$ such that for every $k \in \{1,\ldots,N\}$, $\mathsf{cost}(\mathbb{C}_k) \leq 8 \times \mathsf{optcost}(k)$ (for every set and distance metric).

References

[DL05] Sanjoy Dasgupta and Philip M. Long. Performance guarantees for hierarchical clustering. *J. Comput. Syst. Sci.*, 70(4):555–569, 2005.