Max Planck Institute for Intelligent Systems, Tübingen, Germany

Foundations of Comparison-Based Hierarchical Clustering

Debarghya Ghoshdastidar, **Michaël Perrot**, and Ulrike von Luxburg

Objective: Study hierarchical clustering when only similarity comparisons are available, that is without features nor explicit similarities.

Comparison-Based Machine Learning

Humans are bad at giving unbiased, quantitative information. Better at giving *relative information*.
Example: The left vehicles are *more similar* to each other than the right vehicles.

Given an unknown similarity function w, the corresponding quadruplet is

$$w\left(\text{SUV left}, \text{SUV right}\right) \geqslant w\left(\text{Sport car}, \text{Tractor}\right).$$

Challenging problem: No features (coordinates), not even distances! Given a list of quadruplets, can we solve standard machine learning tasks such as *clustering*?

Example: Let $\mathcal{X} = \{x_i\}_{i=1}^N$ be a set of N cars. Can we build a dendrogram that reflects their similarities using only a limited set of quadruplets $\mathcal{Q}$?

Existing solutions:

– *Embedding based methods*: Retrieve a Euclidean representation of the objects that respects the quadruplets, then use standard machine learning methods.

– *Direct methods*: Design learning algorithms that directly handle the quadruplets to solve a specific task.

Obtaining the comparisons:

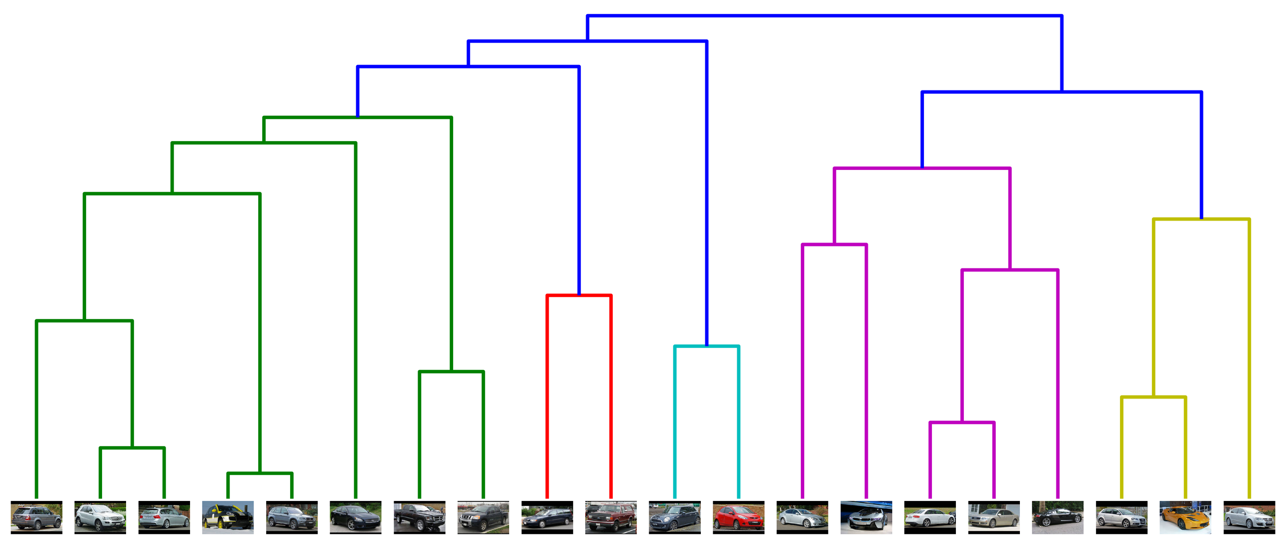
– *Actively*: quadruplets chosen by the algorithm.

– *Passively*: quadruplets given to the algorithm with no way to make new queries.

Contributions:
We propose new algorithms for hierarchical clustering that *directly* use quadruplets.
We derive *sufficient conditions* that guarantee exact recovery of a planted model.

Hierarchical Clustering

Goal: Construct a dendrogram that reflects the similarities between the objects.



Idea: Iteratively group clusters together using a linkage function. Given two clusters G and G':

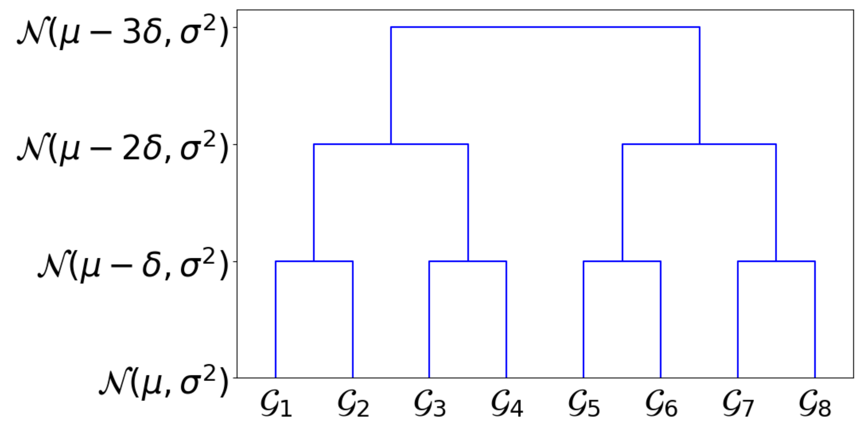
– Single Linkage (SL): $W(G,G') = \max_{x_i \in G, x_j \in G'} w_{ij}$,

– Complete Linkage (CL): $W(G,G') = \min_{x_i \in G, x_j \in G'} w_{ij}$,

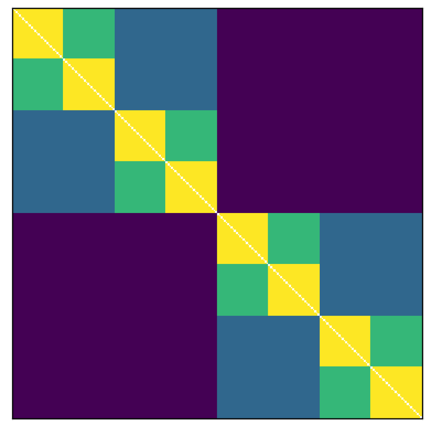
– Average Linkage (AL): $W(G,G') = \sum_{x_i \in G, x_j \in G'} \frac{w_{ij}}{|G||G'|}$.

Planted Model

Summary: We study hierarchical clustering under a noisy hierarchical block matrix. The model complexity is controlled by the size N_0 of the pure clusters, by the number of levels L in the hierarchy, and by the signal to noise ratio $\frac{\delta}{\sigma}$.



Hierarchical structure.



Expected similarities.

Single Linkage (SL) and Complete Linkage (CL)

Summary: Single and Complete linkage are inherently comparison-based. They require at least $\Omega\left(N^2\right)$ and at most $\mathcal{O}\left(N^2 \ln N\right)$ active quadruplets. If the signal to noise ratio grows with the number of examples they can recover the hierarchy. This is tight for Single linkage.

Theorem (Exact recovery of planted hierarchy by SL and CL).

– *If $\frac{\delta}{\sigma} \geqslant \Omega\left(\sqrt{\ln\left(N\right)}\right)$, then SL and CL recover the planted hierarchy with high probability.*

– *If $\frac{\delta}{\sigma} \leqslant \mathcal{O}\left(\sqrt{\ln\left(\frac{N}{2^L}\right)}\right)$ with large $\frac{N}{2^L}$, then SL fails to recover the hierarchy with probability $\frac{1}{2}$.*

Quadruplets Kernel Average Linkage (4K–AL)

Summary: We use the quadruplets to derive a proxy for the similarities between the examples and obtain better guarantees than SL and CL in terms of recovery of the planted model.

Kernel function: Two similar objects should behave similarly with respect to any third object.

– **Active comparisons:** Let $w_{i_0 j_0}$ be a reference similarity and $\mathcal{S}$ be a set of landmarks:

$$K_{ij} = \sum_{k \in \mathcal{S}\setminus\{i,j\}} \left(\mathbb{I}_{\left(w_{ik} > w_{i_0 j_0}\right)} - \mathbb{I}_{\left(w_{ik} < w_{i_0 j_0}\right)}\right)\left(\mathbb{I}_{\left(w_{jk} > w_{i_0 j_0}\right)} - \mathbb{I}_{\left(w_{jk} < w_{i_0 j_0}\right)}\right).$$

– **Passive comparisons:** Use all the similarities as references and all the examples as landmarks:

$$K_{ij} = \sum_{k,l=1,k<l}^{N} \sum_{r=1}^{N} \left(\mathbb{I}_{(i,r,k,l)\in\mathcal{Q}} - \mathbb{I}_{(k,l,i,r)\in\mathcal{Q}}\right)\left(\mathbb{I}_{(j,r,k,l)\in\mathcal{Q}} - \mathbb{I}_{(k,l,j,r)\in\mathcal{Q}}\right)$$

Guarantees: With a constant signal to noise ratio and a sufficient number of comparisons, 4K-AL recovers the planted hierarchy.

Theorem (Exact recovery of planted hierarchy by 4K–AL).

– ***Active comparisons:*** *With $L = \mathcal{O}(1)$, $N_0 \geqslant \Omega\left(\sqrt{N}\right)$, and $\frac{\delta}{\sigma}$ constant, 4K–AL exactly recovers the planted hierarchy with high probability using only $\mathcal{O}\left(N \ln N\right)$ quadruplets.*

– ***Passive comparisons:*** *With $L = \mathcal{O}(1)$, $N_0 \geqslant \Omega\left(\sqrt{N}\right)$, and $\frac{\delta}{\sigma}$ constant, 4K–AL exactly recovers the planted hierarchy with high probability using $\mathcal{O}\left(N^{7/2} \ln N\right)$ quadruplets.*

Quadruplets-Based Average Linkage (4–AL)

Summary: We use passive comparisons to define a cluster-level similarity function. If sufficiently large initial clusters are provided, 4–AL obtains better guarantees than 4K–AL.

Cluster-level similarity: Clusters G_1, G_2 are more similar to each other than G_3, G_4 if their objects are, on average, more similar to each other than the objects of G_3 and G_4:

$$\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_3, G_4\right) = \sum_{x_i \in G_1} \sum_{x_j \in G_2} \sum_{x_k \in G_3} \sum_{x_l \in G_4} \frac{\mathbb{I}_{(i,j,k,l)\in\mathcal{Q}} - \mathbb{I}_{(k,l,i,j)\in\mathcal{Q}}}{|G_1|\,|G_2|\,|G_3|\,|G_4|}.$$

Averaging over all cluster pairs gives rise to the following linkage function:

$$W\left(G_p, G_q\right) = \sum_{r,s=1, r\neq s}^{K} \frac{\mathbb{W}_{\mathcal{Q}}\left(G_p, G_q \| G_r, G_s\right)}{K(K-1)}.$$

Guarantees: With sufficiently large initial clusters, a constant signal to noise ratio, and a sufficient number of comparisons, 4–AL exactly recovers the planted hierarchy.

Theorem (Exact recovery of planted hierarchy by 4–AL).

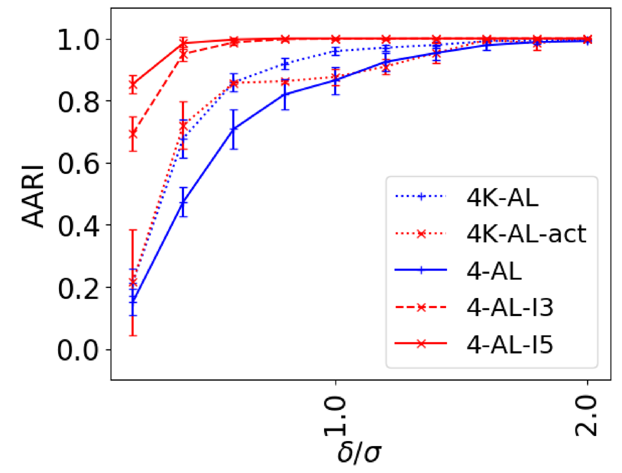
– *With $L = \mathcal{O}(1)$, $\frac{\delta}{\sigma}$ constant, and an initial partition of the examples in pure clusters of sizes in the range $[m, 2m]$ for some $m \leqslant \frac{1}{2}N_0$, 4–AL exactly recovers the planted hierarchy with high probability using $\mathcal{O}\left(\frac{N^4 \ln N}{m}\right)$ passive quadruplets.*

– *With $L = \mathcal{O}(1)$, $\frac{\delta}{\sigma}$ constant, and $\Omega\left(N_0\right)$-sized initial clusters, 4–AL exactly recovers the planted hierarchy with high probability using only $\mathcal{O}\left(N^3 \ln N\right)$ passive quadruplets.*

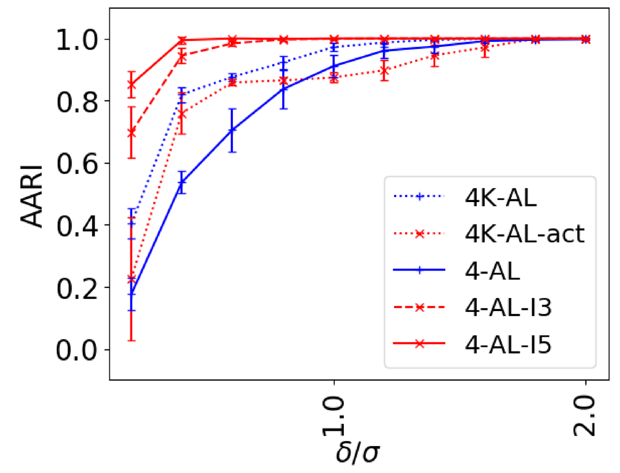
Experiments: Planted Model

Summary: We empirically verify our theoretical findings: SL and CL only recover the hierarchy for large signal to noise ratios while 4k–AL and 4–AL exactly recover the hierarchy for smaller signal to noise ratios.

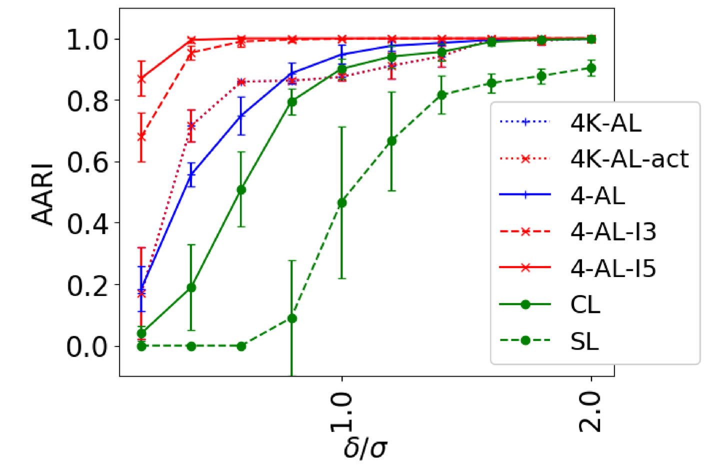
Evaluation: We use the Average Adjusted Rand Index (AARI, higher is better).



Proportion of quadruplets $p = 0.01$.



Proportion of quadruplets $p = 0.1$.



Proportion of quadruplets $p = 1$.

Code available online!