Active Clustering : Robust and Efficient Hierarchical Clustering

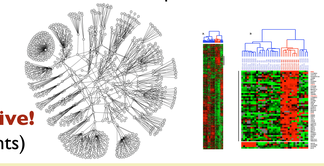
Brian Eriksson[1] Gautam Dasarathy[2] Aarti Singh[3] Robert Nowak[2]
dasarathy@wisc.edu

1. Department of Computer Science, Boston University 2. Department of Electrical Engineering, UW-Madison 3. Department of Machine Learning, Carnegie Mellon University

1. Introduction

Hierarchical Clustering from pairwise similarities is an important tool in a wide range of problems.

- Gene behavior from microarrays
- Network topology discovery
- Community structure in social networks etc.,



Obtaining similarities might be expensive!
(resource intensive experiments, expert judgments)

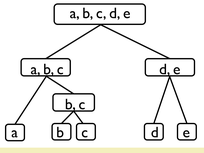
While usual hierarchical clustering techniques require ***O(N²)*** pairwise similarities, we propose an algorithm which, under certain assumptions, requires only ***O(N log N)*** *actively selected* similarities. Further, we propose another algorithm which requires only **O(N log² N)** pairwise similarities *even if there is some "noise" in the similarity values.*

2. Problem Setup

$\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ is a set of N objects and $S = \{s_{ij}\}$ is the set of $\binom{N}{2}$ pairwise similarities

Any subset $\mathcal{C}$ of $\mathcal{X}$ is called a cluster and a collection of clusters $\mathcal{T}$ is called a **hierarchical clustering** if every pair of clusters are either disjoint or one is contained in the other. Assume that there is a true hierarchical clustering $\mathcal{T}$ associated with $\mathcal{X}$

w.l.o.g hierarchical clustering ⇒ binary tree



Goal : Recover *T* using *as few pairwise similarities* as possible

This requires some notion of consistency. This is our other assumption about the problem.

Tight Clustering Condition

The triple $(\mathcal{X}, \mathcal{T}, S)$ is said to satisfy the Tight Clustering (TC) condition, if for any three objects $x_i, x_j, x_k \in \mathcal{X}$ such that $x_i, x_j \in \mathcal{C}$ and $x_k \notin \mathcal{C}$ for some $\mathcal{C} \in \mathcal{T}$, the following holds

$$s_{ij} > \max\{s_{ik}, s_{jk}\}$$

Before we proceed, does randomly selecting similarities work?

- We show that picking similarities at random does not generally work
- In fact, simple arguments show that ***O(N²/m)*** similarities are required to reconstruct a cluster of size *m*.

3. Efficient Hierarchical Clustering

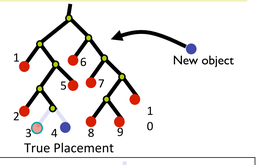
This method **adaptively** selects the *most informative* pairwise similarities in order to reconstruct the hierarchical clustering *T*.
It was inspired by an algorithm proposed in [1] for learning graphical models of binary random variables and is based on the following ***"outlier" test***

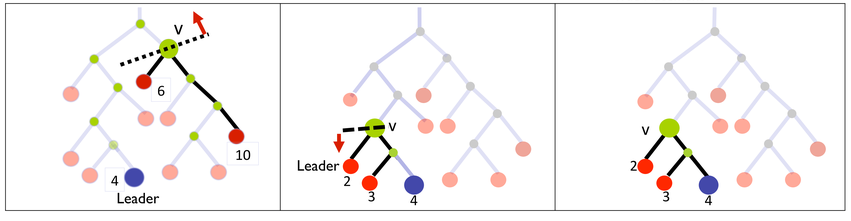
$$\texttt{outlier}(x_i, x_j, x_k) = x_i \text{ if } s_{jk} > \max\{s_{ik}, s_{ij}\}, \text{ similarly for } x_j \text{ and } x_k$$

TC condition ⇒ *outlier*(● , ● , ●) = ● if and only if 

Inserting a new object into`` a tree with *i* leaves

- Pick an internal *v* node with ≈ *i*/2 objects as descendants
- Find two leaves x_k *and* x_j whose common ancestor is *v*
- Find *outlier*(x_k, x_j, *v*) and discard a portion of the tree
- Proceed till there are only two leaves left and insert using a final *outlier* test.





Theorem 1: If *(X,T,S)* satisfies the TC condition, then *T* can be recovered using no more than **3N $\log_{1.5}$ N** adaptively selected pairwise similarities.

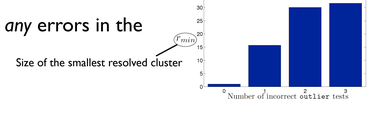
Efficient Hierarchical Clustering : Performance

Topology	Size (*N*)	n_{agg}	$n_{outlier}$	$n_{outlier}/n_{agg}$
Balanced Binary Tree	128	8,128	876	10.78%
	256	32,640	2,206	6.21%
	512	130,816	5,561	4.25%
Synthetic Internet	768	294,528	8,490	2.88%

$\mathbf{n_{agg}}$: Number of pairwise similarities required by Hierarchical Agglomerative Clustering
$\mathbf{n_{outlier}}$: Number of pairwise similarities required by our method

Fragility

The algorithm is greedy. Therefore it is highly sensitive to *any* errors in the outlier test.



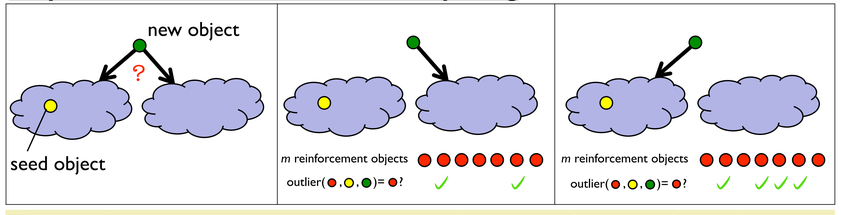
4. Robust and Efficient Hierarchical Clustering

Model :
Assume that a subset of the similarities S_C are consistent wrt to the TC condition on *T*. Further, assume that any given similarity fails to be consistent with probability $q < ½$.

Strategy : Top-Down Recursive Splitting
To split a cluster *C* reliably, we roughly do the following

- Pick a "seed" object and pick *m* "reinforcement" objects.
- For every other object in *C*, measure how similar it is with respect to the seed object using outlier tests on the reinforcement objects.
- If similar, put in the same cluster as the seed. Else put in the other cluster. Repeat on each resulting cluster till no cluster has more than *m* objects

Simplified Illustration of Robust Cluster Splitting



Theorem 2: Under certain assumptions on *q* and η (which controls how unbalanced clusters can be), for any *0*<δ<*1*, there exists a constant $k_0(\eta,\delta,q)$ such that our algorithm recovers all clusters in *T* such that

- The size of the cluster is bigger than $2m = k_0 \log(8N/\delta)$
- All clusters that contain this cluster should not be too imbalanced

Further the procedure requires *no more than* **O(N log²N)** pairwise similarities

Synthetic Experiments (N=512)

q	Agglomerative Clustering Δ-Entropy	Agglomerative Clustering r_{min}	Robust (m=40) Δ-Entropy	Robust (m=40) r_{min}	Robust (m=80) Δ-Entropy	Robust (m=80) r_{min}
0.05	0.3666	460.8	1.0178	6.8	1.0178	7.2
0.15	0.0899	512	1.0161	16	1.0161	15.2
0.25	0.0133	512	0.09360	384	1.0119	57.6

Gene Microarray Experiments

Dataset [2]	Agglo. Δ-Entropy	Robust (m=10) Δ-Entropy	Robust (m=10) n_{robust}/n_{agg}	Robust (m=80) Δ-Entropy	Robust (m=80) n_{robust}/n_{agg}
Gene (N=512)	0.1417	0.1912	27%	0.2035	53%
Gene (N=1024)	0.0761	0.1325	18%	0.1703	36%

Δ-Entropy quantifies how good the clustering is. **Bigger the better.**
$\mathbf{r_{min}}$ is the smallest cluster that can be resolved. **Smaller the better.**

5. References

[1] J. Pearl and M. Tarsi, "Structuring Causal Trees," in *Journal of Complexity*, vol. 2, 1986, pp. 60–77.
[2] J. DeRisi, V. Iyer, and P. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," in *Science*, vol. 278, October 1997, pp. 680–686.