Functorial Clustering via Simplicial Complexes

Dan Shiebler

University of Oxford

Objectives

1. Define the maximal and single linkage algorithms in terms of the finite fuzzy singular set functor.
2. Reformulate existing results on hierarchical overlapping clustering algorithms [CGHS16] in terms of functors that factor through a category of simplicial complexes.
3. Introduce a functorial strategy for using a finite clustering to partition an infinite space.

Extrapolation

In practice we often need to extrapolate a clustering to out-of-sample points. Say we have a flat clustering functor C, a not-necessarily-finite uber-metric space $(\mathbf{X}, d_{\mathbf{X}})$, and some finite $X \subset \mathbf{X}$. We want to produce a covering of $(\mathbf{X}, d_{\mathbf{X}})$ by grouping the points in $\mathbf{X} - X$ into the sets in $C(X, d_{\mathbf{X}})$. Intuitively, we want to do this in a way such that if the points $x' \in X$ and $x \in \mathbf{X} - X$ would be placed into the same cluster if we ran C on $X \cup \{x\}$, then they will share a set in this covering of $(\mathbf{X}, d_{\mathbf{X}})$.

To do this, first define a functor $C_{X \cup \{x\}}$ that maps uber-metric spaces of the form $(X \cup \{x\}, d_{\mathbf{X}})$ to the maximal cover that is refined by $C(X, d_{\mathbf{X}}) \cup \{\{x\}\}$ and refines $C(X \cup \{x\}, d_{\mathbf{X}})$. The cover $C_{X \cup \{x\}}(X \cup \{x\}, d_{\mathbf{X}})$ is identical to $C(X, d_{\mathbf{X}})$, except some of the sets in this cover will also contain the point x. Intuitively, $C_{X \cup \{x\}}$ assigns each $x \in \mathbf{X} - X$ to the sets in $C(X, d_{\mathbf{X}})$ that contain the points in X that share a cluster with x in $C(X \cup \{x\}, d_{\mathbf{X}})$. In order to stitch together each of these assignments into a cover of $\mathbf{X}$, we can simply take the colimit of the functor $C_{X \cup \{x\}}$. Intuitively, this colimit is a cover of $\mathbf{X}$ that is refined by $C(X, d_{\mathbf{X}}) \cup \{\{x_i\} \mid x_i \in \mathbf{X} - X\}$.

Flat Clustering Definitions

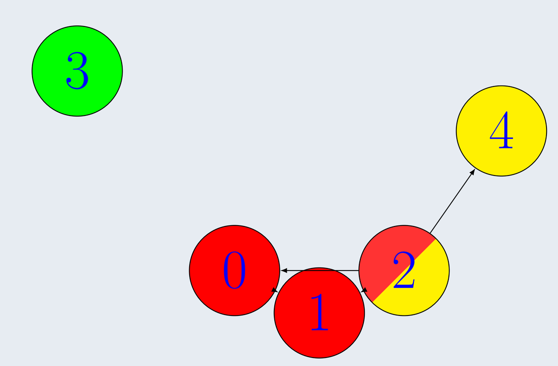
- In the category **UMet** objects are finite uber-metric spaces and morphisms are non-expansive maps.
- Given a set X, a **non-nested flag cover** $\mathcal{C}_X$ of X is a cover of X such that: (1) if $A, B \in \mathcal{C}_X$ and $A \subseteq B$, then $A = B$, (2) the simplicial complex with vertices corresponding to the elements of X and faces all finite subsets of the sets in $\mathcal{C}_X$ is a flag complex.
- The category **Cov** has tuples $(X, \mathcal{C}_X)$ as objects where $\mathcal{C}_X$ is a non-nested flag cover of the finite set X. The morphisms between $(X, \mathcal{C}_X)$ and $(Y, \mathcal{C}_Y)$ are refinement-preserving functions.
- A **flat clustering functor** is a functor $C : \mathbf{UMet} \to \mathbf{Cov}$ that is the identity on the underlying set.

Hierarchical Clustering Definitions

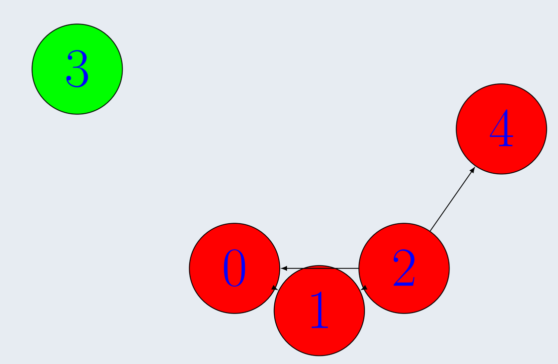
- A **fibered fuzzy simplicial complex** is a functor $F_X : I^{op} \to \mathbf{SCpx}$ such that for any morphism $a \leq a'$ in I^{op}, the simplicial map $F_X(a \leq a')$ acts as the identity on 0-simplices.
- **Fuzzy non-nested covers** are functors $F_X : I^{op} \to \mathbf{Cov}$ such that $S_{fl} \circ F_X$ is a fibered fuzzy simplicial complex. The category of fuzzy non-nested covers and natural transformations is **FCov**.
- The functors $S_{fl} : \mathbf{Cov} \to \mathbf{SCpx}$ and $Flag : \mathbf{SCpx} \to \mathbf{Cov}$ are adjoint functors that map between fibered fuzzy simplicial complexes and fuzzy non-nested covers
- A **hierarchical clustering functor** is a functor $H : \mathbf{UMet} \to \mathbf{FCov}$ such that for $a \in (0, 1]$, $H(-)(a) : \mathbf{UMet} \to \mathbf{Cov}$ is a flat clustering functor.

Single and Maximal Linkage

- The functor $Pair : \mathbf{UMet} \to \mathbf{FSCpx}$ sends the uber-metric space $(X, d_X) \in \mathbf{UMet}$ to the fibered fuzzy simplicial complex $F_X : I^{op} \to \mathbf{FSCpx}$ where for $a \in (0, 1]$, $F_X(a)$ is a simplicial complex whose set of 0-simplices is X and whose 1-simplices are the pairs $\{x_1, x_2\} \subseteq X$ such that $d_X(x_1, x_2) \leq -\log(a)$. $F_X(a)$ has no n-simplices for $n > 1$.
- Define the **finite singular set functor** as $FinSing = (S_{fl} \circ -) \circ (Flag \circ -) \circ Pair$
- **Maximal Linkage** $\mathcal{ML} = (Flag \circ -) \circ FinSing$
  - The points x_1, x_2 lie in the same cluster with strength at least a if the largest pairwise distance between them is no larger than $-\log(a)$.



- **Single Linkage** $\mathcal{SL} = (\pi_0 \circ -) \circ FinSing$
  - The points $x_1, x_2 \in X$ lie in the same cluster with strength at least a if there exists a sequence of points $x_1, x_i, x_{i+1}, ..., x_{i+n}, x_2$ such that $d(x_j, x_{j+1}) \leq -\log(a)$



Universality of Single/Maximal Linkage

Intuitively, single linkage and maximal linkage clustering lie on two ends of a spectrum of clustering refinement. Any other non-trivial hierarchical clustering functor lies between them. Formally, we can make the following claim, which is inspired by Theorem 8 in Culbertson et al [CGHS16]:

Suppose $H : \mathbf{UMet} \to \mathbf{FCov}$ is a non-trivial hierarchical clustering functor such that for all $a \in (0, 1]$, the functor $H(-)(a) : \mathbf{UMet} \to \mathbf{Cov}$ has clustering parameter $\delta_{H,a}$ and define $W_H(a) = e^{-\delta_{H,a}}$. Then there exist natural transformations with inclusion maps as components from $\mathcal{ML}(-)(W_H(-))$ to H and from H to $\mathcal{SL}(-)(W_H(-))$.

References

[CGHS16] Jared Culbertson, Dan P Guralnik, Jakob Hansen, and Peter F Stiller. Consistency constraints for overlapping data clustering. *arXiv preprint arXiv:1608.04331*, 2016.

Acknowledgements

I would like to thank my advisors, Jeremy Gibbons and Cezar Ionescu, for their help and support.

Contact Information

- Web: danshiebler.com
- Email: daniel.shiebler@kellogg.ox.ac.uk