Some ideas for formalizing clustering schemes

Gunnar Carlsson
Mathematics Department
Stanford University
California, CA 94305
gunnar@math.stanford.edu

Facundo Mémoli
Mathematics Department
Stanford University
California, CA 94305
memoli@math.stanford.edu

Abstract

Despite being one of the most commonly used tools for unsupervised exploratory data analisys and despite its and extensive literature very little is known about the theoretical foundations of clustering methods. We have been working on various mathematical approaches which allow the extension of earlier results in this area based on ideas from topology and metric geometry. We will give an overview at the workshop.

1 Clustering

Desirable properties of clustering algorithms come from practicioners who have intuitive notions of what is a good clustering: they know it when they see it. These is of course not satifactory and a theoretical understanding needs to be developed. We want to argue that one thing this intuition reflects is the fact that density needs to be incorporated in the clustering procedures. Single linkage clustering, a procedure that enjoys several nice theoretical properties, is notorious for its insensitivity to density, which is manifested in the so called *chaining effect*. Other methods such as average linkage, complete linkage and k-means share the property that they exhibit some sort of sensitivivity to density, but are unstable in a sense which can be made theoretically precise and are therefore not well supported by theory. We believe that this disconnect between theory and practice should not exist. In particular, in [1] we have constructed a framework that incorporates density via the use of 2-dimensional persistence ideas.

Kleinberg's impossibility theorem [2] illustrates an important feature of clustering methods which is that fixing a scale parameter is a very awkward thing. Statisticians developed hierarchical clustering, a style of clustering that provides a summary of the behaviour of clusters at all choices of the scale parameter at once. This changes the notion of what a clustering scheme is and one of the tasks we face is to understand what is the analogue of Kleinberg's theorem for this new class of *hierarchical* clustering schemes. Dendrograms, the output of such methods, offer a good summary when there is just one scale parameter in the hierarchical notion of clustering. In order to incorporate density we need more than just one parameter and the challenge is to find useful summaries, similar to dendrograms, for the situation when there are 2 or more parameters.

We are trying to close the gap between the theoretical side and the wishes of practicioner in two different ways. One is the explicit inclusion of density into a multiparameter clustering scheme [1]. The other is by constructing a theory that permits a good deal of variety in the clustering schemes, and in particular, includes analogues of the clique clustering mehtods familiar in network and graph problems [3, 4].

2 Overview of some of our results

Standard clustering methods take as input a finite metric space (X, d) and output a partition of X. Kleinberg [2] discussed this situation in an axiomatic way and identified a set of reasonable properties of standard clustering schemes, namely, scale invariance, richness and consistency. He then proved, in the spirit of Arrow's impossibility theorem, that no clustering scheme satisfying these conditions simultaneously can exist. In the same spirit as Kleinberg's theorem, we prove that in the context of HC methods, one obtains existence and uniqueness instead of non-existence.

Hierarchical clustering: formulation Given a finite metric space (X, d), a hierarchical clustering method f returns a nested family of partitions, or **dendrogram** (a.k.a. persistent set) of X:

$$f(X,d) \in \mathcal{D}(X) \;=\; \{(X,\theta) |\, \theta : [0,\infty) \to \mathcal{P}(X)\} \text{ such that}$$

(1) $\theta(0) = \{\{x_1\}, \ldots, \{x_n\}\}$; **(2)** there exists t_0 s.t. $\theta(t)$ is the *single block partition* for all $t \geqslant t_0$; **(3)** if $r \leqslant s$ then $\theta(r)$ *refines* $\theta(s)$; and **(4)** for all r there exists $\varepsilon > 0$ s.t. $\theta(r) = \theta(t)$ for $t \in [r, r+\varepsilon]$.

Following [5], we represent dendrograms (= weighted rooted trees) as *ultrametric* spaces: a metric space (X, u) is an ultrametric space if and only if for all $x, x', x'' \in X$, $\max(u(x,x'), u(x',x'')) \geqslant u(x,x'')$. For $n \in \mathbb{N}$ let $\mathcal{X}_n$ (resp. $\mathcal{U}_n$) denote the set of all metric spaces (resp. ultra-metric spaces) with n points. Let $\mathcal{X} = \sqcup_{n\geqslant 1}\mathcal{X}_n$ denote set of all finite metric spaces and $\mathcal{U} = \sqcup_{n\geqslant 1}\mathcal{U}_n$ all finite ultrametric spaces. Then, a hierarchical clustering method can be regarded as a map $T : \mathcal{X} \to \mathcal{U}$ s.t. $\mathcal{X}_n \ni (X,d) \mapsto (X,u) \in \mathcal{U}_n$, $n \in \mathbb{N}$. There is a canonical construction: Let $T^* : \mathcal{X} \to \mathcal{U}$ be given by $(X,d) \mapsto (X, u^*)$ where

$$u^*(x,x') := \min \left\{ \max_{i=0,\ldots,\ell-1} d(x_i, x_{i+1}) |\, x = x_0, \ldots, x_\ell = x' \right\}.$$

This construction yields exactly *single linkage clustering*, [6]. For $X \in \mathcal{X}$ let $\mathrm{sep}(X,d) := \min_{x\neq x'} d(x,x')$. We have the following *characterization theorem*:

Theorem 2.1 ([5]) *Let T be a clustering method s.t.*

1. $T\left(\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right)\right) = \left(\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right)\right)$ *for all* $\delta > 0$.
2. *Whenever $X, Y \in \mathcal{X}$ and $\phi : X \to Y$ are such that $d_X(x,x') \geqslant d_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$, then it also holds that*
$$u_X(x,x') \geqslant u_Y(\phi(x), \phi(x'))$$
for all $x, x' \in X$, where $T(X, d_X) = (X, u_X)$ and $T(Y, d_Y) = (Y, u_Y)$.
3. *For all $(X,d) \in \mathcal{X}$,*
$$u(x,x') \geqslant \mathit{sep}(X,d) \text{ for all } x \neq x' \in X$$
where $T(X,d) = (X,u)$.

Then $T = T^*$, i.e., T is single linkage HC.

Remark 2.1 *It is interesting to consider the case when one requires ϕ to be 1 to 1 on points. In this case, a much wider class of hierarchical schmemes becomes possible including for example a certain version of* clique clustering. *The restriction on the nature of ϕ would be called restriction of* functoriality *by a mathematician. The classification question of clustering methods that arises becomes mathematically interesting and we are currently exploring it [4].*

Metric stability of T^* We also obtain the Proposition and Theorem below asserting metric stability and asymptotic consistency of the method T^*. We use the notion of Gromov-Hausdorff distance between metric spaces, [7]. The Gromov-Hausdorff distance $d_{\mathcal{GH}}(X,Y)$ between compact metric spaces (X, d_X) and (Y, d_Y) is defined to be the infimal $\varepsilon > 0$ s.t. there exists a metric d on $X \sqcup Y$ with $d_{|_{X\times X}} = d_X$ and $d_{|_{Y\times Y}} = d_Y$ for which the Hausdorff distance between X and Y (as subsets of $(X \sqcup Y, d)$) is less than ε. This distance is a natural choice for comparing dendrograms as well (when viewed as ultrametric spaces), see Figure 1 and [5].

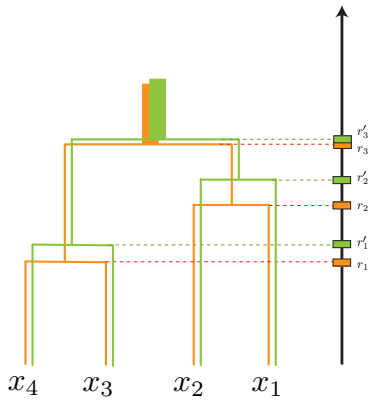

Figure 1: Two different dendrograms (X, α) and (X, β) over the same underlying set $X = \{x_1, x_2, x_3, x_4\}$. Let α be the dendrogram represented in orange and β be the one in green. The condition that $d_{\mathcal{GH}}((X, \alpha), (X, \beta)) \leqslant \varepsilon/2$ is equivalent to the horizontal dotted lines corresponding to r_i and r'_i $(i = 1, 2, 3)$ being within ε of eachother.

Proposition 2.1 ([5]) *For any finite metric spaces (X, d_X) and (Y, d_Y)*

$$d_{\mathcal{GH}}((X, d_X), (Y, d_Y)) \geqslant d_{\mathcal{GH}}(T^*(X, d_X), T^*(Y, d_Y)).$$

Fix a finite set X. For a symmetric function $W : X \times X \to \mathbb{R}^+$ let $\mathcal{L}(W)$ denote the maximal metric on X less than of equal to W:

$$\mathcal{L}(W)(x, x') = \min \left\{ \sum_{i=0}^{m} W(x_i, x_{i+1}) |\, x = x_0, \ldots, x_m = x \right\} \text{ for } x, x' \in X.$$

Theorem 2.2 ([5]) *Assume (Z, d_Z) is a compact metric space. Let X and X' be any two finite sets of points sampled from Z and $r, r' > 0$ such that $Z \subset \cup_{x \in X} B(x, r)$ and $Z \subset \cup_{x' \in X'} B(x', r')$. Let $d_X = d_{Z|_{X \times X}}$ and $d_{X'} = d_{Z|_{X' \times X'}}$. Let $T^*(X, d_X) = (X, u_X)$ and $T^*(X', d_{X'}) = (X', u_{X'})$. Then one has*

1. *(Finite Stability) $d_{\mathcal{GH}}((X, u_X), (X', u_{X'})) \leqslant (r + r')$.*
2. *(Convergence/consistency) Assume in addition that $Z = \sqcup_{\alpha \in A} Z_\alpha$ where A is a finite index set and Z_α are compact, disjoint and path-connected sets. Let (A, d_A) be the finite metric space with underlying set A and metric given by $d_A := \mathcal{L}(W)$ where $W(\alpha, \alpha') := \min_{z \in Z_\alpha, z' \in Z_{\alpha'}} d_Z(z, z')$ for $\alpha, \alpha' \in A$. Let $T^*(A, d_A) = (A, u_A)$. Then, as $r \to 0$ one has $d_{\mathcal{GH}}((X, u_X), (A, u_A)) \to 0$.*

2.1 Two-parameter clustering

Despite its wide applicability, there is an unresolved issue regarding the difference of single linkage (SL) clustering in contrast to average and complete linkage (AL and CL) clustering. In spite of the fact that SL enjoys nice theoretical properties, practicioners have found the *chaining effect* inherent to SL to be unacceptable. On the other hand, it is known that AL and CL are *unstable* in a precise sense, but practical applications are usually done with one of these two methods, and this seems to yield reasonable results. With regard to the chaining effect, it is well understood that one of the shortcomings of SL is it insensitivity to *density*. In this direction, a classical result of Hartigan [8] proves that SL is not *consistent* in the sense that it is unable to recover modes of an underlying density in $\mathbb{R}^d$. In [1] propose an extension of HC, called *multiparameter hierarchical clustering methods* that tries to remedy this situation. The input to the method we propose is a triple (X, d, f), where (X, d) is a finite metric space and $f : X \to \mathbb{R}$ is a function defined on the data X, which could be a density estimate or could represent some other type of information.

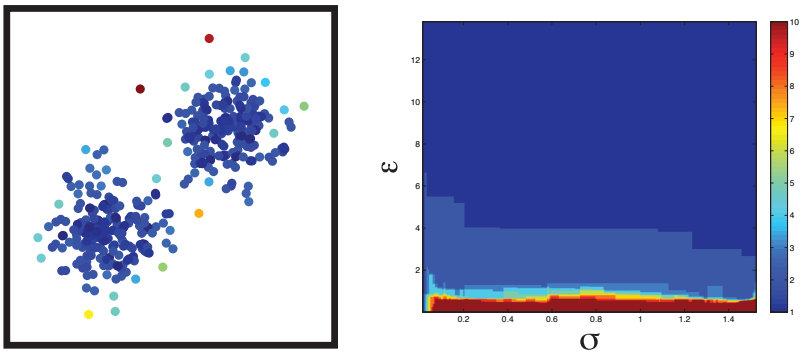

Figure 2: In the output of the method of [1] we track two parameters, as opposed to the dendrograms of HC that track just one. In our formulation the input is (X, d, f), ε tracks the metric/scale information and σ tracks the information contained in the sub-level sets of the function f. The example in this figure shows a dataset exhibiting two modes. The colors of the points in reflect the inverse of some density estimate. On the right, we show the output. The colorbar indicates the number of clusters observed for each value of the pair (ε, σ). Notice that for large values of ε, one sees only one component, whereas for a large range of epsilons two components are detected.

The output of our method is more general than dendrograms, see Figure 2. Our construction is motivated by the methods of *persistent topology* [9], the Reeb graph and Cluster Trees [10]. In [1] we obtained a characterization theorem and establish a stability results which are similar to Theorems 2.1 and 2.2. The computation of the two-dimensional dendrograms yields problems that can be solved in polynomial time [11].

References

[1] Gunnar Carlsson and Facundo Mémoli. Multiparameter clustering methods. In *IFCS 2009*, 2009.

[2] Jon M. Kleinberg. An impossibility theorem for clustering. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 446–453. MIT Press, 2002.

[3] G. Carlsson and F. Mémoli. Persistent Clustering and a Theorem of J. Kleinberg. *ArXiv e-prints*, August 2008.

[4] Gunnar Carlsson and Facundo Mémoli. Classifying clustering schemes. Technical report, 2009. In preparation.

[5] Gunnar Carlsson and Facundo Mémoli. Characterization, stability and convergence of hierarchical clustering algorithms. Technical report, 2009.

[6] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice Hall Advanced Reference Series. Prentice Hall Inc., Englewood Cliffs, NJ, 1988.

[7] D. Burago, Y. Burago, and S. Ivanov. *A Course in Metric Geometry*, volume 33 of *AMS Graduate Studies in Math.* American Mathematical Society, 2001.

[8] J. A. Hartigan. Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.*, 76(374):388–394, 1981.

[9] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Proc. 41st Ann. IEEE Sympos. Found Comput. Sci.*, pages 454–463, 2000.

[10] Werner Stuetzle. Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, 20(1):25–47, 2003.

[11] Gunnar Carlsson, Gurjeet Singh, and Afra Zomorodian. Computing multidimensional persistence. *ArXiv e-prints*, July 2009.