Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Functorial hierarchical clustering with overlaps

Jared Culbertson [a], Dan P. Guralnik [b,*], Peter F. Stiller [c]

[a] Sensors Directorate, Air Force Research Laboratory, 2241 Avionics Circle, Building 620 Wright–Patterson Air Force Base, OH 45433-7302, USA
[b] Electrical & Systems Engineering Department, University of Pennsylvania, 200 S. 33rd st., 203 Moore Bldg. Philadelphia, PA 19104-6314, USA
[c] Department of Mathematics, MS3368, Texas A& M University College Station, TX 77843-3368, USA

ARTICLE INFO

Article history:
Received 21 October 2016
Received in revised form 18 August 2017
Accepted 12 October 2017
Available online xxxx

Keywords:
Hierarchical clustering
Clustering with overlaps
Functorial clustering
Clustering domain
Injective envelope
Non-expansive map
Sieving functor
Cut metric
Tree metric
A-space

ABSTRACT

This work draws inspiration from three important sources of research on dissimilarity-based clustering and intertwines those three threads into a consistent principled functorial theory of clustering. Those three are the overlapping clustering of Jardine and Sibson, the functorial approach of Carlsson and Mémoli to partition-based clustering, and the Isbell/Dress school's study of injective envelopes. Carlsson and Mémoli introduce the idea of viewing clustering methods as functors from a category of metric spaces to a category of clusters, with functoriality subsuming many desirable properties. Our first series of results extends their theory of functorial clustering schemes to methods that allow overlapping clusters in the spirit of Jardine and Sibson. This obviates some of the unpleasant effects of chaining that occur, for example with single-linkage clustering. We prove an equivalence between these general overlapping clustering functors and projections of weight spaces to what we term clustering domains, by focusing on the order structure determined by the morphisms. As a specific application of this machinery, we are able to prove that there are no functorial projections to cut metrics, or even to tree metrics. Finally, although we focus less on the construction of clustering methods (clustering domains) derived from injective envelopes, we lay out some preliminary results, that hopefully will give a feel for how the third leg of the stool comes into play.

1. Introduction

Problems surrounding data clustering have been studied extensively over the last forty years. Clustering stands as an important tool for analyzing and revealing the often hidden structure in data (and in today's big data) coming from fields as diverse as biology, psychology, machine learning, sociology, image understanding, and chemistry. Among the earliest systematic treatments of clustering theory was that of Jardine and Sibson in 1971 [33]. They laid out important desiderata for overlapping clustering methods and provided a relatively efficient algorithm for their so-called B_k clustering which allowed overlapping clusters with no more than $k-1$ points in any overlap. Since then, there have been several distinct directions of research in clustering theory, with only modest linkage between the methods of researchers pursuing different paths.

The classical work of Jardine and Sibson was followed by other similarly comprehensive works such as Everitt [24]. Further theoretical work on these mostly classical methods was also done by Kleinberg [34] and Carlsson and Mémoli [9,10]. Kleinberg in particular showed the incompatibility of a relatively simple set of desirable axioms for any partition based clustering method. Carlsson and Mémoli in turn introduced categorical language into partition based clustering and showed

[*] Corresponding author.
E-mail addresses: jared.culbertson@us.af.mil (J. Culbertson), guraldan@seas.upenn.edu (D.P. Guralnik), stiller@math.tamu.edu (P.F. Stiller).

https://doi.org/10.1016/j.dam.2017.10.015
0166-218X/© 2017 Elsevier B.V. All rights reserved.

that single-linkage was (up to a scaling) the only functorial method satisfying all their axioms (which included the notions of representability and excisiveness) in the category of finite metric spaces with non-expansive maps.

In another direction, work on computing phylogenetic trees inspired a seminal paper by Bandelt and Dress [3] on split decompositions of metrics. This line of research was continued with investigations into split systems and cut points of injective envelopes of metric spaces. Representative papers include [21] and [22]. While not explicitly clustering methods, these methods are quite similar in spirit to stratified/hierarchical clustering schemes. In this genre, we might also add the classification of the injective envelopes of six-point metric spaces by Sturmfels and Yu [40]. Bandelt and Dress have also had a large influence on another field as a result of their work on weak hierarchies [2,4]. This led to work by Diatta, Bertrand, Barthélemy, Brucker, and others on indexed set systems (see, e.g., [5,6,14]). Another interesting development in this area is the work by Janowitz on ordinal clustering [32].

Recently, with the emergence of the new field of topological data analysis (TDA), work has been done on topologically-based clustering methods. This includes the Mapper algorithm by Singh, Mémoli, and Carlsson [39], as well as work on persistence-based methods [11] and Reeb graphs [28].

Meanwhile, most users of clustering methods default either to a classical linkage-based clustering method (such as single-linkage or complete linkage) or to more geometrically-based methods like k-means. Unfortunately, the wide array of clustering theories has had little impact on the actual practice of clustering. Simply put, the gap between theory and efficient practice has been hard to bridge.

In this paper we draw inspiration from three of the sources mentioned above, and strive to intertwine those three threads into a consistent principled functorial theory of dissimilarity-based clustering. Those three are the overlapping clustering of Jardine and Sibson, the functorial approach of Carlsson and Mémoli, and the Dress school approach to clustering, via injective envelopes, which were independently discovered by Isbell and Dress. This paper intends to fuse these approaches. Our starting point is the paper of Carlsson and Mémoli [10] which introduces the idea of viewing clustering methods as functors from a category of metric spaces to a category of clusters. Many desirable properties of a clustering method are subsumed in functoriality when the morphisms are properly chosen. Here the relevant morphisms under which the particular method is functorial can be viewed as giving restrictions on the allowable data processing operations—restrictions that impose consistency constraints across related data sets. One of our first goals is to extend their theory of functorial clustering schemes to methods that allow overlapping clusters in the spirit of Jardine and Sibson, and in so doing obviate some of the unpleasant effects of chaining occurring in some linkage-based methods. (See [9], Remark 16, where Carlsson and Mémoli discuss the chaining effects in single-linkage clustering and propose an alternate solution based on explicitly considering density.) Rather than relying on chaining to overcome certain technical problems, we accept overlapping clusters. This leads to a much richer set of possible clustering algorithms.

Finally, although in this paper we focus less on the construction of clustering methods (clustering domains) derived from injective envelopes, we do in the final section lay out some preliminaries, that hopefully will enable the reader to get a feel for how the geometry of injective envelopes comes into play. In addition, lest the reader think we are all theory and no practice, we mention our ongoing algorithmic work on efficient implementations of some of these clustering schemes, along the lines of what has already been done for q-metrics by Segarra et al. [37] and for dithered maximal linkage clustering by Gama et al. [25,26].

1.1. Weight categories

Definition 1. Let **Weight** be the category of **finite sets with weights**, whose objects have the form (X, u) with X a finite non-empty set and u a symmetric non-negative map $u : X \times X \to \mathbb{R}$, $(x, y) \mapsto u_{xy}$ satisfying $u_{xx} = 0$ for all $x \in X$. A morphism $f : (X, u) \to (Y, v)$ is a set map $f : X \to Y$ such that $v_{f(x)f(x')} \leq u_{xx'}$; these will be referred to as *non-expansive maps*.

For a fixed finite set X, we can define a local order structure by pointwise dominance on the full subcategory $\mathbf{Weight}_X$ of **Weight** consisting of weights on X. In order to simplify notation, when the underlying set X is fixed we will often refer to the object $(X, u) \in \mathbf{Weight}_X$ only by the weight function u and state that $u \in \mathbf{Weight}_X$. The set of objects of $\mathbf{Weight}_X$ is a partially ordered set (poset) with the ordering given by

$$u \leq v \text{ if } u_{xy} \leq v_{xy} \text{ for all } x, y \in X.$$

It will be convenient to denote, for any subset U of the objects of $\mathbf{Weight}_X$,

$$U\downarrow := \left\{ w \in \mathbf{Weight}_X \,\middle|\, \exists u \in U \; w \leq u \right\}$$

and in the case of the singleton set $\{u\}$, we will often write $u\downarrow$ for $\{u\}\downarrow$. For any subcategory $\mathbf{C}$ of **Weight**, we will use the analogous notation $\mathbf{C}_X$ for $\mathbf{C} \cap \mathbf{Weight}_X$, which on objects will be the intersection of the objects of $\mathbf{C}$ and $\mathbf{Weight}_X$ with the morphisms of $\mathbf{C}$. Also, for every map of finite sets $f : X \to Y$ and $w \in \mathbf{Weight}_Y$ we define $f^*(w) \in \mathbf{Weight}_X$ to be the pullback of the weight w to the set X, more explicitly, $f^*(w)_{xy} := w_{f(x)f(y)}$. This notation allows another perspective on morphisms in **Weight**: the map of finite sets $f : X \to Y$ induces a morphism $(X, u) \to (Y, v)$ in **Weight** if and only if $f^*(v) \leq u$.

Please cite this article in press as: Culbertson–Guralnik–Stiller, Functorial hierarchical clustering with overlaps, Discrete Applied Mathematics (2017), https://doi.org/10.1016/j.dam.2017.10.015.