A Characterization of Linkage-Based Clustering: An Extended Abstract

Margareta Ackerman, Shai Ben-David and David Loker
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1
{mackerma,shai,dloker}@cs.uwaterloo.ca

1 Introduction

There are a wide variety of clustering algorithms that, when run on the same data, often produce very different clusterings. Yet there is no principled method to guide the selection of a clustering algorithm. The choice of an appropriate clustering is, of course, task dependent. As such, we must rely on domain knowledge. The challenge is to communicate such knowledge between the domain expert and the algorithm designer. One approach to providing guidance to clustering users in the selection of a clustering algorithm is to identify important properties that a user may want an algorithm to satisfy, and determine which algorithms satisfy each of these properties. Clustering users can then utilize prior knowledge to determine the properties that make sense for their application.

Ultimately, there would be a sufficiently rich set of properties that would provide detailed enough guidelines for a wide variety of clustering users. For a property to be useful, a user needs to be able to easily determine the desirability of the property. Such a description of clustering algorithms would yield principled guidelines for clustering algorithm selection by answering a series of simple questions. Bosagh Zadeh and Ben-David [1] make progress in this direction by providing a set of abstract properties that characterize single linkage. In this work, we give another result in the same direction by characterizing a family of clustering algorithms. These are initial steps toward the ambitious program of developing broad guidelines for clustering algorithm selection.

Linkage-based clustering is one of the most commonly-used and widely-studied clustering paradigms. We provide a surprisingly simple set of properties that uniquely identify linkage-based clustering algorithms. Our characterization highlights how linkage-based algorithms compare to other clustering algorithms.

Combining previously proposed properties with our newly proposed ones, we show how these properties partition the space of commonly-used clustering algorithms. Specifically, we show which of these properties are satisfied by common linkage-based, centroid-based, and spectral clustering algorithms. We hope that this analysis, as well as our characterization of linkage-based clustering, will provide useful guidelines for users in selecting clustering algorithms.

2 Notation

A *distance function* is a function $d : X \times X \to R^+$, such that d is symmetric and $d(x, x) = 0$ for all $x \in X$. Let $P(X)$ denote all subsets of data set X.

For clustering C, let $|C|$ denote the number of clusters in C. For $x, y \in X$ and clustering C of X, we write $x \sim_C y$ if x and y belong to the same cluster in C and $x \not\sim_C y$, otherwise.

A *k-clustering* $C = \{c_1, c_2, \ldots, c_k\}$ of data set X is a set of k disjoint subsets of X, such that $\bigcup_i c_i = X$. A *clustering* of X is a k-clustering of X for some $1 \leqslant k \leqslant |X|$.

3 Defining linkage-based Clustering

There is little variety in the literature on the definition of linkage-based clustering, and the differences that do exist are syntactic in nature. A linkage-based algorithm begins by placing every element of a data set into its own cluster, and then repeatedly merges clusters. What distinguishes among different linkage-based algorithms is the mechanism used to determine which clusters to merge at every step. There are a number of ways to describe the mechanism by which the algorithm decides which clusters to merge. Such a mechanism is often described as a "linkage function," which takes two clusters as input and outputs a real number (see, for example, [2] and [4]). Everitte et al. [3] refer to a linkage function as an "inter-object distance." Although neither are rigorously defined, they represent the same concept. We expect that the inter-object distance depends on the distances between the individual points. Further, we expect that the measure depends *only* on the pairwise distances between the points in the data set, and not on factors such as the labels of the points. As such, the inter-object distance can be thought of as an "extension operator" of the original distance function.

Our focus here is on the precise definition of the extension operator. Previous definitions of linkage-based clustering vary slightly and are often informal on this point.

Definition 1. *An* extension operator $\wedge$ *of* d *over* X *is a function* $\hat{d} : P(X)^2 \to R^+$.

Since the extension operator "sees" all of d, it can make decisions based on distances outside the two input clusters. That is, $\hat{d}(A, B)$ may depend on distances outside of $A \cup B$. Moreover, $\hat{d}(A, B)$ may depend on the point labels. To overcome these shortcomings, we present a more strict version of an extension operator.

Definition 2 (Local extension operator)**.** *A* local extension operator *is an extension operator* $\wedge$ *so that for all data sets* A, B, C, D *and distance functions* d_1, d_2*,*

$$\hat{d_1}(A, B) = \hat{d_2}(C, D),$$

whenever there exists a distance-preserving isomorphism $\phi : A \cup B \to C \cup D$ *where* $d_2(\phi(x), \phi(y)) = d_1(x, y)$ *for all* $x, y \in A \cup B$*,* $C = \{\phi(a) \mid a \in A\}$ *and* $D = \{\phi(b) \mid b \in B\}$.

Given a local extension operator, $\hat{d}(A, B)$ depends only on the distances in $A \cup B$.

Definition 3 (linkage-based clustering function)**.** *A clustering function* F *is* linkage-based *if there exists a local extension operator* $\wedge$ *so that*

- $F(X, d, |X|) = \{\{x\} \mid x \in X\}$
- *For* $1 \leqslant k < |X|$*,* $F(X, d, k)$ *is constructed by merging the two clusters in* $F(X, d, k + 1)$ *that minimize the value of* $\hat{d}$*. Formally,*

  $$F(X, d, k) = \{c \mid c \in F(X, d, k + 1), c \neq c_i, c \neq c_j\} \cup \{c_i \cup c_j\},$$

  such that $\{c_i, c_j\} = argmin_{\{c_i, c_j\} \subseteq F(X,d,(k+1))} \hat{d}(c_i, c_j)$.

Observe that single-linkage, average-linkage, and complete-linkage are linkage-based clustering functions.

In the event that a distance function that depends on the entire data is desirable, one can simply use an extension operator that is not local. We call such a clustering function is *data-dependent linkage-based*.

4 Properties

4.1 Hierarchical clustering

Hierarchical clustering is a widely used class of clustering algorithms. Here we give a concise formalization on what makes a clustering algorithm hierarchical.

Definition 4 (Refinement)**.** *A clustering* C *of* X *is a refinement of clustering* C' *of* X *if every cluster* $c'_i \in C'$ *is the union of some clusters in* C.

Definition 5 (Hierarchical)**.** *A clustering function is* hierarchical *if for every* $1 \leqslant k \leqslant k' \leqslant |X|$*,* $F(X, d, k')$ *is a refinement of* $F(X, d, k)$.

4.2 Locality

We now introduce locality as a property of clustering algorithms. In our main result, we show how local extension operators relate to local clustering algorithms.

Definition 6 (local)**.** *A clustering function F is* local *if for any clustering $C \subseteq F(X, d, k)$,*

$$F(X', d/X', |C|) = C$$

where $X' = \bigcup_{c \in C} c$. That is, X' consists of the points in clusters of C.

To better understand locality, consider two runs of a clustering algorithm. In the first run, the algorithm is called on some data set X and returns a k-clustering C. We then select some clusters $c_1, c_2, \ldots, c_{k'}$ of C, and run the clustering algorithm on the points that the selected clusters consist of, namely, $c_1 \cup c_2 \cup \ldots \cup c_{k'}$ asking for k' clusters. If the algorithm is local, then on the second run of the algorithm it will output $\{c_1, c_2, \ldots, c_{k'}\}$.

While locality is an intuitive property, it is not satisfied by all clustering functions. For instance, some spectral clustering functions are not local.

4.3 Consistency

Consistency, introduced by Kleinberg [5], requires that the output of a clustering function, which takes the number of clusters as input, be invariant to shrinking within-cluster distances, and stretching between-cluster distances.

Definition 7 (consistency)**.** *A clustering function F is* consistent *if $F(X, d_X, k) = F(X, d'_X, k)$ whenever*

- *$d'_X(x, y) \leqslant d_X(x, y)$ if $x \sim_{F(X,d_X,k)} y$, and*
- *$d'_X(x, y) \geqslant d_X(x, y)$ if $x \not\sim_{F(X,d_X,k)} y$.*

for all X,d_X, and $1 \leqslant k \leqslant |X|$.

We introduce two weak variations of consistency.

Definition 8 (outer-consistency)**.** *A clustering function F is* outer-consistent *if $F(X, d_X, k) = F(X, d'_X, k)$ whenever $d'_X(x, y) \geqslant d_X(x, y)$ if $x \not\sim_{F(X,d_X,k)} y$ and $d'_X(x, y) = d_X(x, y)$ if $x \sim_{F(X,d_X,k)} y$, for all X,d_X, and $1 \leqslant k \leqslant |X|$.*

Definition 9 (inner-consistency)**.** *A clustering function F is* inner-consistent *if $F(X, d_X, k) = F(X, d'_X, k)$ whenever $d'_X(x, y) \leqslant d_X(x, y)$ if $x \sim_{F(X,d_X,k)} y$ and $d'_X(x, y) = d_X(x, y)$ if $x \not\sim_{F(X,d_X,k)} y$, for all X,d_X, and $1 \leqslant k \leqslant |X|$.*

Clearly, consistency implies outer-consistency and inner-consistency.

We show in Section 6 that many common clustering algorithms, including the most common linkage-based algorithms, satisfy outer-consistency. On the other hand, there are many clustering algorithms, including maximum-linkage and average-linkage, that fail inner-consistency.

5 Characterizing linkage-based Clustering

The following is our main result, which specifies the properties that uniquely identify linkage-based clustering functions.

Theorem 1. *An outer-consistent clustering function is linkage-based if and only if it is hierarchical and local.*

Data-dependent linkage-based clustering functions are linkage-based clustering functions whose extension operators may depend on the entire distance function. To characterize these clustering functions, we no longer require locality.

Theorem 2. *An outer-consistent clustering function is data-dependent linkage-based if and only if it is hierarchical.*

6 Taxonomy of clustering

We show how the properties discussed above partition the space of commonly-used clustering algorithms. In particular, we analyze which properties are satisfied by linkage-based (single linkage, average linkage and complete linkage), spectral clustering (ratio-cut and normalized cut), and centroid-based algorithms (k-means and k-median). For example, while locality is an intuitive property, it is not satisfied by spectral clustering algorithms. We also show that complete linkage and average linkage both fail inner-consistency, and therefore also fail consistency. Below is a table outlining our results.

Clustering Algorithm	Outer-consistency	Inner-consistency	Locality	Hierarchical	Order-invariance
Single Linkage	✓	✓	✓	✓	✓
Average Linkage	✓	*X*	✓	✓	*X*
Complete Linkage	✓	*X*	✓	✓	✓
K-means	✓	*X*	✓	*X*	*X*
K-median	✓	*X*	✓	*X*	*X*
Ratio-cut	✓	✓	*X*	*X*	*X*
Normalized-cut	✓	*X*	*X*	*X*	*X*

References

[1] Reza Bosagh Zadeh and Shai Ben-David. "A Uniqueness Theorem for Clustering." The 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI 09), 2009.

[2] Chris Ding and Xiaofeng He. "Cluster Aggregate Inequality and Multi-level Hierarchical Clustering." Knowledge Discovery in Databases (PKDD), 2005. LNCS, Springer Berlin / Heidelberg, V. 3721, pp. 71-83

[3] Brain Everitt, Sabine Landau, and Morven Leese. "Cluster Analysis." Fourth Edition. Oxford University Press. 2001.

[4] Derek Greene, Gerard Cagney, Nevan Krogan, and Padraig Cunningham. "Ensemble non-negative matrix factorization methods for clustering proteinprotein interactions." Bioinformatics Vol. 24 no. 15 2008, pages 17221728

[5] Jon Kleinberg. "An Impossibility Theorem for Clustering." Advances in Neural Information Processing Systems (NIPS) 15, 2002.