A Characterization of Linkage-Based Algorithms

David Loker

Joint work with

Margareta Ackerman and Shai Ben-David

Motivation

- There are a wide variety of clustering algorithms, which often produce very different clusterings.
- How should a user decide which algorithm to use for a given application?

Our approach

- Identify properties that distinguish between the results of different clustering paradigms
- The properties should be:
  1) Intuitive and “user-friendly”
  2) Useful for classifying clustering algorithms
- Clustering users can utilize prior knowledge to determine which properties make sense for their application
- Then use these properties to sort out clustering algorithms

Previous work

- Kleinberg proposes abstract properties (“Axioms”) of clustering functions (NIPS, 2002)
- Bosagh Zadeh and Ben-David provide a set of properties that characterize single linkage clustering (UAI, 2009)

Our contributions

- Propose a couple of properties that uniquely indentify linkage-based clustering algorithms
- Construct a taxonomy of clustering algorithms based on the properties

Outline

- Define linkage-based clustering
- Our new clustering properties
- Main result
- Sketch of proof
- A taxonomy of common clustering algorithms using clustering properties
- Conclusions

Formal setup

For a finite domain set X, a *distance function* d over the members of X.

A Clustering Function F *maps*

Input: (X,d) and $k>0$

to

Output: a k-partition (clustering) of X

Linkage-based algorithm: An informal definition

- Start with the clustering of singletons
- Merge the closest pair of clusters
- Repeat until only k clusters remain.

?

Informal definition of between-cluster distance
“An extension of the between-point distance that applies to subsets of the domain”

- The definition of between-cluster distances is what distinguishes between linkage-based algorithms.

Outline

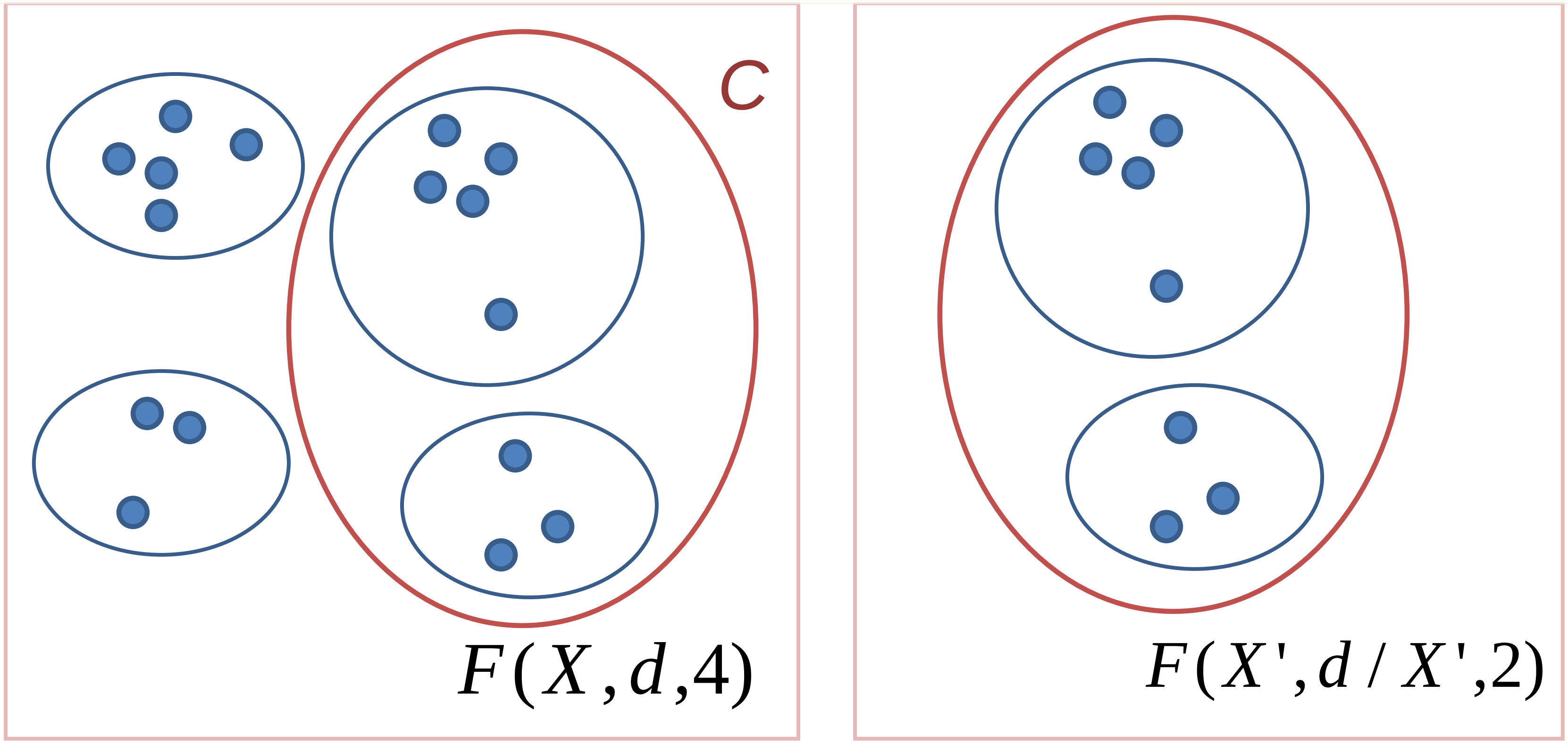
- Define linkage-based clustering
- Our new clustering properties
- Main result
- Sketch of proof
- A taxonomy of common clustering algorithms using our properties
- Conclusions

Hierarchical clustering

- A clustering C is a *refinement* of clustering C' if every cluster c' in C' is a union of some clusters in C.
- A clustering function is *hierarchical* if for every $1 \leq k \leq k' \leq |X|$,

$F(X,d,k')$ is a refinement of $F(X,d,k)$.

Locality



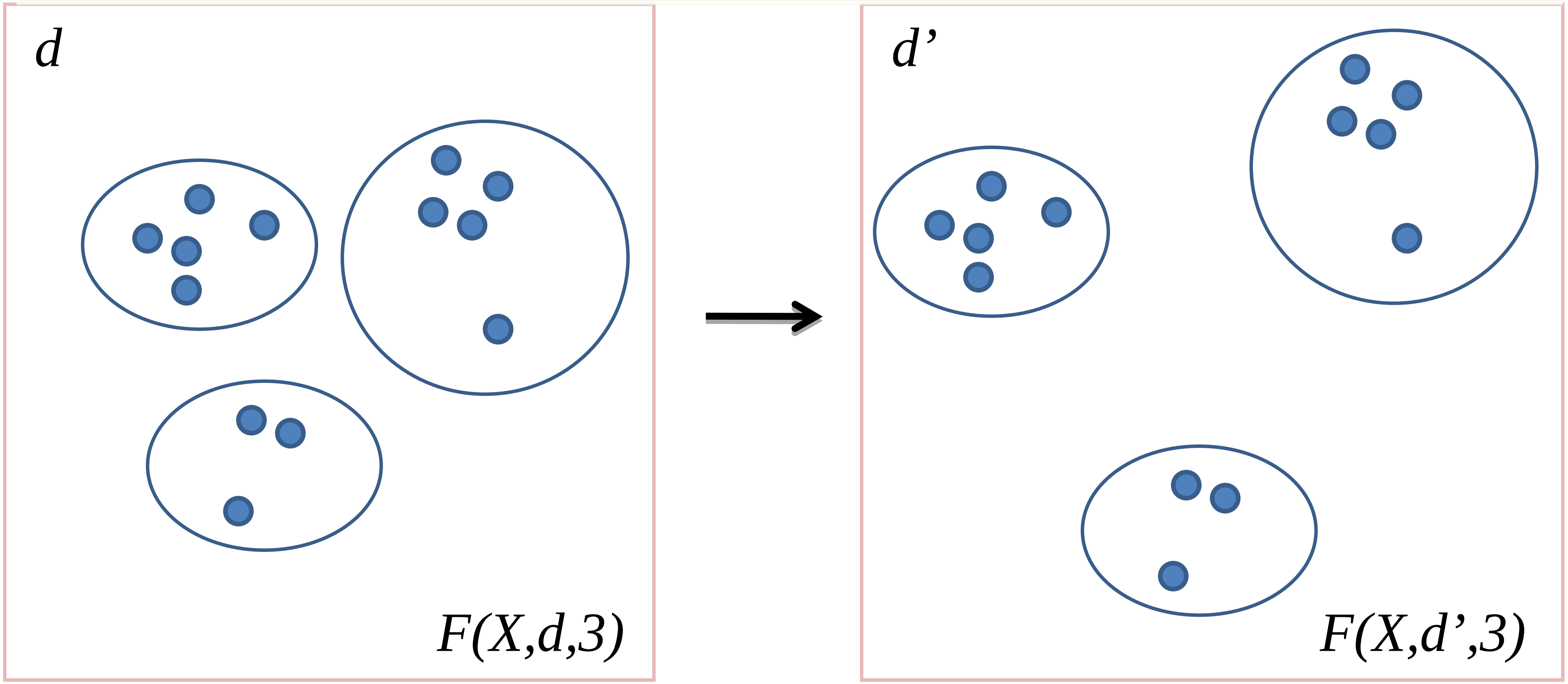
F is *local* if for any $C \subseteq F(X,d,k)$,

$$C = F(X' = \bigcup_{c \in C} c, d/X', |C|)$$

Which paradigms satisfy locality ?

- Many clustering algorithms are local
  - K-means
  - K-median
  - Single-linkage
  - Average-linkage
  - Complete-linkage
- Notably, not all clustering algorithms satisfy locality
  - Ratio cut
  - Normalized cut

Outer Consistency
Based on Kleinberg, 2002.



If d' equals d, except for increasing between-cluster distances, then $F(X,d,k)=F(X,d',k)$.

Which paradigms satisfy outer-consistency?

Many clustering algorithms satisfy outer-consistency

- K-means
- K-median
- Single-linkage
- Average-linkage
- Complete-linkage
- Ratio cut
- Normalized cut

Outline

- Define linkage-based clustering
- Our new clustering properties
- Main result
- Sketch of proof
- A taxonomy of common clustering algorithms using our properties
- Conclusions

Our main result

Theorem:

An outer-consistent clustering function is linkage based iff it is hierarchical and local.

Easy direction of proof

Any linkage based clustering function is hierarchical and local

The proof is quite straight-forward.

Interesting direction of proof

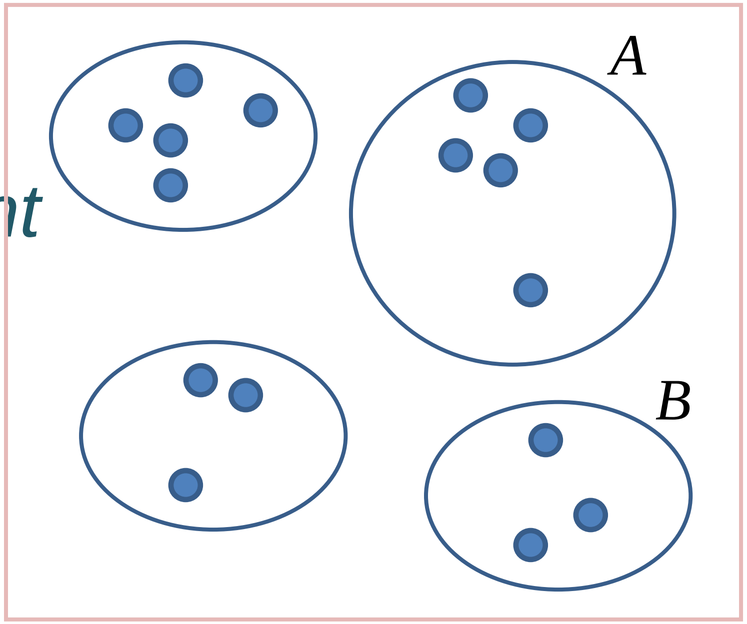
If *F* is outer-consistent, local, and hierarchical then *F* is linkage-based

To prove this direction we first need to formalize linkage-based clustering, by formally defining between-cluster distance

What do we expect from the between-cluster distance?

1) Extends the point-wise distance: $d: X^2 \rightarrow R^+$

2) The distance between *subsets* A and B is *independent* of data outside of these two clusters



3) Usually assumes no ties

Sketch of proof

Recall direction:

If F is outer-consistent, local, and hierarchical then F is linkage-based.

Goal:

Given d over X, find a between-cluster distance $\bar{d} : P(X)^2 \rightarrow R^+$, such that the following algorithm outputs $F(X,d,k)$:

– Start with the clustering of singletons
– Merge the two clusters that minimize $\bar{d}$
– Repeat until k clusters remain

Sketch of proof (continued…)

- For $A, B, C, D \subseteq X$, $(A,B) < (C,D)$ if there exists a dataset $X' \subseteq X$ s.t. when applying the clustering function to X', A and B are merged before C and D
- Prove that $<$ is an ordering (anti-symmetric and transitive)
- Use the ordering to construct $\bar{d}$

Outline

- Define linkage-based clustering
- Our new clustering properties
- Main result
- Sketch of proof
- A taxonomy of common clustering algorithms using our properties
- Conclusions

Taxonomy of clustering algorithms

	local	Outer consistent	Inner consistent	hierarchical	Order Invariant
Single linkage	✓	✓	✓	✓	✓
Average linkage	✓	✓	✗	✓	✗
Complete linkage	✓	✓	✗	✓	✓
K-means	✓	✓	✗	✗	✗
K-median	✓	✓	✗	✗	✗
Ratio-cut	✗	✓	✓	✗	✗
Normalized-cut	✗	✓	✗	✗	✗

Taxonomy of clustering algorithms

	local	Outer consistent	Inner consistent	hierarchical	Order Invariant
Single linkage	✓	✓	✓	✓	✓
Average linkage	✓	✓	✗	✓	✗
Complete linkage	✓	✓	✗	✓	✓
K-means	✓	✓	✗	✗	✗
K-median	✓	✓	✗	✗	✗
Ratio-cut	✗	✓	✓	✗	✗
Normalized-cut	✗	✓	✗	✗	✗

Taxonomy of clustering algorithms

	local	Outer consistent	Inner consistent	hierarchical	Order Invariant
Single linkage	✓	✓	✓	✓	✓
Average linkage	✓	✓	✗	✓	✗
Complete linkage	✓	✓	✗	✓	✓
K-means	✓	✓	✗	✗	✗
K-median	✓	✓	✗	✗	✗
Ratio-cut	✗	✓	✓	✗	✗
Normalized-cut	✗	✓	✗	✗	✗

Taxonomy of clustering algorithms

	local	Outer consistent	Inner consistent	hierarchical	Order Invariant
Single linkage	✓	✓	✓	✓	✓
Average linkage	✓	✓	✘	✓	✘
Complete linkage	✓	✓	✘	✓	✓
K-means	✓	✓	✘	✘	✘
K-median	✓	✓	✘	✘	✘
Ratio-cut	✘	✓	✓	✘	✘
Normalized-cut	✘	✓	✘	✘	✘

Taxonomy of clustering algorithms

	local	Outer consistent	Inner consistent	hierarchical	Order Invariant
Single linkage	✓	✓	✓	✓	✓
Average linkage	✓	✓	✗	✓	✗
Complete linkage	✓	✓	✗	✓	✓
K-means	✓	✓	✗	✗	✗
K-median	✓	✓	✗	✗	✗
Ratio-cut	✗	✓	✓	✗	✗
Normalized-cut	✗	✓	✗	✗	✗

Conclusions

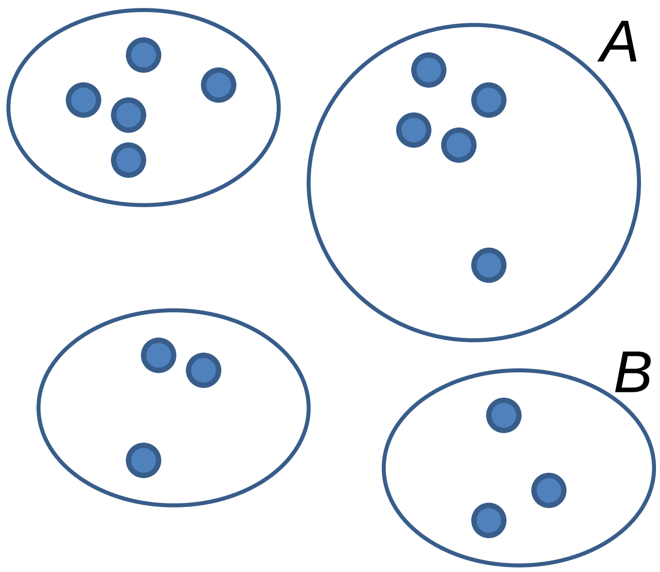
- We introduced new properties of clustering algorithms.
- We showed that an outer-consistent clustering algorithm is linkage-based iff it is hierarchical and local.
- We classified common clustering algorithms using these properties.

Defining cluster distance

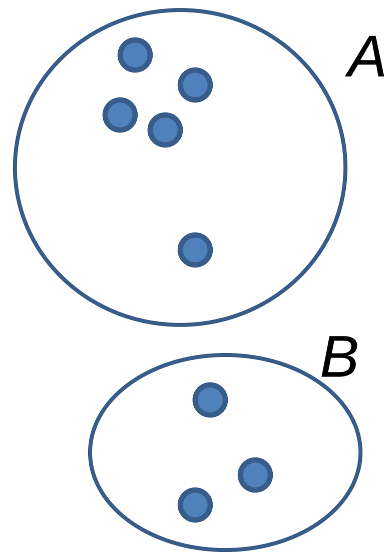
An *extension operator* is an injection

$\overline{d} : P(X)^2 \to R^+$ such that for all A,B

$$\overline{d}(A,B) = \overline{d\,/(A,B)}(A,B)$$



$\overline{d}(A,B) = 5$



$\overline{d\,/(A,B)}(A,B) = 5$