Attempts to Axiomatize Clustering

Shai Ben-David

University of Waterloo, Canada

NIPS Workshop
December 2005

Workshop Goals

Assuming we agree that theory is needed,

We wish to create a basis for a research community:

- Define/detect concrete open problems.
- Foster common language/ terminology/ classification-of-research-directions, among us.
- Stimulate/ brain-storm.
- Increase awareness of what others are/were doing.

The Theory-Practice Gap

Clustering is one of the most widely used tool for exploratory data analysis.

Social Sciences
Biology
Astronomy
Computer Science
.
.

All apply clustering to gain a first understanding of the structure of large data sets.

Yet, there exist distressingly little theoretical understanding of clustering

The Inherent Obstacle

Clustering is* not *well defined.

There is a wide variety of different clustering tasks, with different (often implicit) measures of quality.

Common Solutions

- ***Consider a restricted set of distributions:***

  Mixtures of Gaussians [Dasgupta ‘99], [Vempala,, ’03], [Kannan et al ‘04], [Achlitopas, McSherry ‘05].

- ***Add structure:***
- ***“Relevant Information” –***
  - Information Bottleneck approach [Tishby, Pereira, Bialek ‘99]
- ***Postulate an Objective Utility/Loss Functions*** –
  - K means
  - Correlation Clustering [Blum, Bansal Chawla]
  - Normalized Cuts [Meila and Shi]

- ***Information Theoretic Objective Functions***:
  - Bregman Divergences [Banerjee, Dhilon, Gosh, Merugu]
  - Rate-distortion [Slonim, Atwal, Tkacik, Bialek]
  - Description length [Cilibrasi-Vitanyi, Myllymaki]

Common Solutions (2)

- ***Fitting Generative Models***
  - Mixture of Gaussians
  - SuperParaMagnetic Clustering [Blatt, Weiseman, Domany]
  - Density Traversal Clustering [Storkey and Griffith]

- ***Focus on specific algorithmic paradigms***
  - Agglomerative techniques (e.g., single linkage) [Hartigan, Stuetzle]
  - Projections based clustering (random/spectral) [Ng, Jordan, Weiss]
  - Spectral-based representations – [Belkin, Niyogi]
  - Unsupervised SVM's [Xu and Schuurmans]

Many more …..

Formalizing the broad notion of clustering – Why?

- Different clustering techniques often lead to qualitatively different results. Which should be used when? (Model selection).
- Evaluating the quality of clustering methods – *currently this is embarrassingly ad hoc.*
- *Distinguishing significant structure from random fata morgana.*
- *Providing performance guarantees for sample-based clustering algorithms.*

- *Much more …*

Some attempts to Axiomatizing Clustering

- Jardine and Sibson (1971),
- Hartigan (1975),
- Jane and Dubes (1981)
- Puzicha-Hofmann-Buhmann (2000)
- Kleinberg (2002)

The Basic Setting

- For a finite domain set S, a *dissimilarity function* (*DF*) is a symmetric mapping $d:SxS \rightarrow R^+$ such that $d(x,y)=0$ iff $x=y$.

- A *clustering function* takes a dissimilarity function on S and returns a partition of S.

We wish to define the properties that distinguish clustering functions (from any other functions that output domain partitions).

Kleinberg's Axioms

- *Scale Invariance*

  $F(\lambda d)=F(d)$ for all d and all non-negative λ.

- *Richness*

  For any finite domain S,

  $\{F(d): d$ *is a DF over* $S\}=\{P:P$ *a partition of* $S\}$

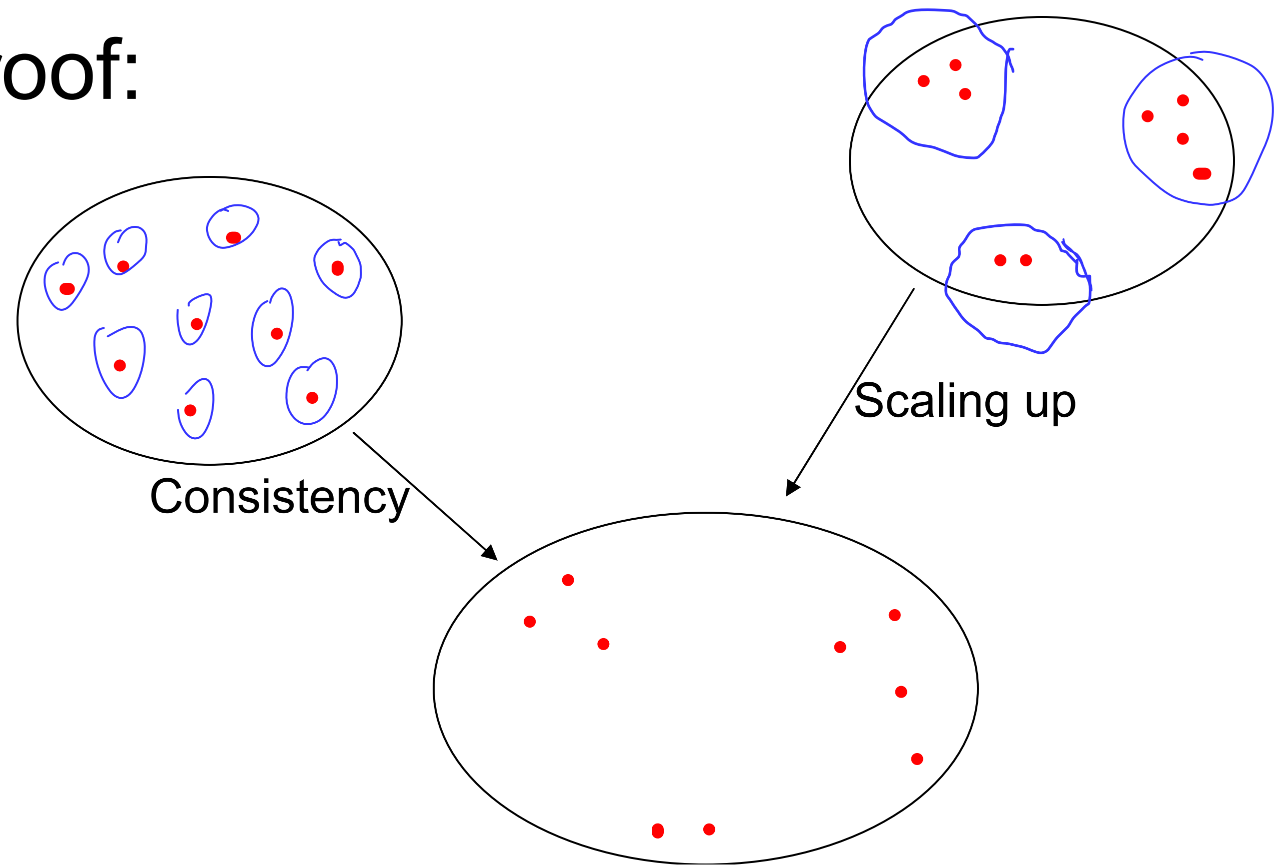
- *Consistency*

  If d' equals d except for shrinking distances within clusters of $F(d)$ or stretching between-cluster distances (w.r.t. $F(d)$), then $F(d)=F(d')$.

Kleinberg's Impossibility result

There exist no clustering function

Proof:



A Different Perspective-
Axioms as a tool for classifying clustering paradigms

- The goal is to generate a variety of axioms (or properties) over a fixed framework, so that different clustering approaches could be classified by the different subsets of axioms they satisfy.

A Different Perspective-
Axioms as a tool for classifying clustering paradigms

- The goal is to generate a variety of axioms (or properties) over a fixed framework, so that different clustering approaches could be classified by the different subsets of axioms they satisfy.

"Axioms" *"Properties"*

	Scale Invariance	Richness	Local Consistency	Full Consistency		
Single Linkage	-	+	+	+		
Center Based	+	+	+	-		
Spectral	+	+	+	-		
MDL	+	+	-			
Rate Distortion	+	+	-			

Ideal Theory

- We would like to have a list of ***simple*** *properties* so that major clustering methods are distinguishable from each other using these properties.
- We would like the *axioms* to be such that *all* methods satisfy *all* of them, and *nothing* that is clearly not a clustering satisfies all of them.

(this is probably too much to hope for).

- *In the remainder of this talk, I would like to discuss some candidate "axioms" and "properties" to get a taste of what this theory-development program may involve*.

Types of Axioms/Properties

- *Richness requirements*

  E.g., relaxations of Kelinberg's richness, e.g.,

  {F(d): d *is a DF over* ***S}={P:P*** *a partition of* ***S*** *into* ***k*** *sets****}***

- *Invariance/Robustness/Stability* requirements.

  E.g., Scale-Invariance, Consistency, robustness to perturbations of ***d*** ("smoothness" of ***F***) or stability w.r.t. sampling of ***S***.

Relaxations of Consistency

- ***Local Consistency*** –

  Let $C_1, \ldots C_k$ be the clusters of $F(d)$.

  For every $\lambda_0 \geq 1$ and positive $\lambda_1, ..\lambda_k \leq 1$, if d' is defined by:

$$d'(a,b) = \begin{cases} \lambda_i d(a,b) & \text{if } a \text{ and } b \text{ are in } C_i \\ \lambda_0 d(a,b) & \text{if } a,b \text{ are not in the same } F(d)\text{-cluster,} \end{cases}$$

then $F(d)=F(d')$.

Is there any known clustering method for which it fails?

(What about Rate Distortion? ..)

Some more structure

- For partitions P_1, P_2 of $\{1, \ldots m\}$ say that P_1 **refines** P_2 if every cluster of P_1 is contained in some cluster of P_2.

- A collection $C=\{P_i\}$ is a **chain** if, for any P, Q, in C, one of them refines the other.

- A collection of partitions is an **antichain,** if no partition there refines another.

- Kleiberg's impossibility result can be rephrased as
  "If F is Scale Invariant and Consistent then its range is an antichain".

Relaxations of Consistency

- ***Refinement Consistency***

  *Same as Consistency (shrink in-cluster, strech between-clusters) but we relax the Consistency requirement "**F(d)=F(d')**" to*

  *"one of **F(d), F(d')** is a refinement of the other".*

- *Note:* A natural version of Single Linkage ("join x,y, iff d(x,y) < λ[max{d(s,t): s,t in X}]") satisfies *this* + *Scale Invariance*+ *Richness*.

  So Kleinberg's impossibility result breaks down.

Should this be an "axiom"?

Is there any common clustering function that fails that?

More on ‘Refinement Consistency’

- *“Minimize Sum of In-Cluster Distances”* satisfies it (as well as *Richness* and *Scale Invariance*).

- *Center-Based clustering* fails to satisfy *Refinement Consistency*

- *This is quite surprising, since they look very much alike.*

$$\sum_{i=1}^{k} \sum_{x,y \in C_i} d^2(x,y) = 2\sum_{i=1}^{k} |C_i| \sum_{x \in C_i} d^2(x,c_i)$$

(Where *d* is Euclidean distance, and c_i the center of mass of C_i)

Hierarchical Clustering

- Hierarchical clustering takes, on top of d, a "coarseness" parameter t.

For any fixed t, $F(t,d)$ is a clustering function.

- We require, for every d:
  - $C_d=\{F(t,d): 0 \le t \le \mathbf{Max}\}$ a chain.
  - $F(0,d)= \{\{x\}: x \in S\}$ and $F(\mathbf{Max},d)=\{S\}$.

Hierarchical versions of axioms

- *Scale Invariance:* For any d, and $\lambda>0$,
  {F(t,d): t} = {F(t, λd):t} (as sets of partitions).

- *Richness:* For any finite domain ***S***,
  {{F(t,d):t}: d *is a DF over* ***S}={C:C*** *a chain of partitions of* ***S*** *(with the needed Min and Max partitions)****}***.

- *Consistency:* If, for some ***t***, ***d'*** is an ***F(t,d)*** -consistent transformation of ***d***, then, for some ***t'***, ***F(t,d)=F(t',d')***

Characterizing Single Linkage

- *Ordinal Clustering* axiom

  If, for all w,x,y,z,

  $d(w,x)<d(y,z)$ iff $d'(w,x)<d'(yz)$

  then $\{F(t,d): t\} = \{F(t,d'):t\}$ (as sets of partitions).

  (note that this implies *Scale Invariance*)

- Hierarchical *Richness* + *Consistency* + *Ordinal Clustering* characterize Single Linkage clustering.

Stability/Robustness axioms

- Relaxing *Invariance* to "*Robustness*"

  Namely, "Small changes in *d* should result in small changes of *f(d)*".

- Statistical setting and *Stability* axioms.

- Axioms as tools for Model Selection.

Sample Based Clustering

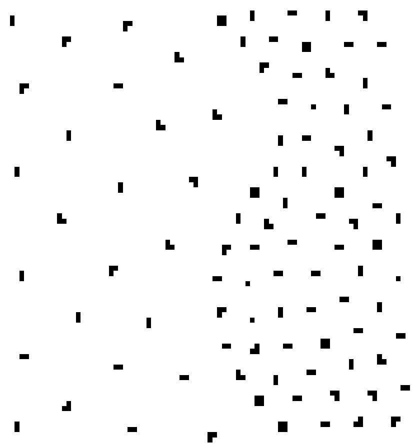
- There is some large, possibly infinite, domain set ***X***.
- An unknown probability distribution ***P*** over ***X*** generates an i. i.d sample, $S \subseteq X$.
- Upon viewing such a sample, a learner wishes to deduce a clustering, as a simple, yet meaningful, description of the distribution.

Stability - basic idea

- *Cluster independent samples of the data.*
- *Compare the resulting clusterings.*
- *Meaningful clusterings should not change much from one independent sample to another.*
- Rational: *To help quantify whether algorithm-generated clusterings reflect properties of the underlying data distribution, rather than being just an artifact of sample randomness.*

Other types of clustering

- Culotta and McCallum's "*Clusterwise Similarity*"
- *Edge-Detection* (advantage to smooth contours)
- *Texture clustering*



-The professors example.

Conclusions and open questions

- There is a place for developing an axiomatic framework for clustering.

- The existing negative results do not rule out the possibility of useful axiomatization.

- We should also develop a system of "clustering properties" for a taxonomy of clustering methods.

- There are many possible routes to take and hidden subtleties in this project.