Clustervision: Visual Supervision of Unsupervised Clustering

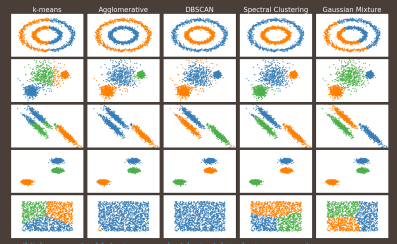
Bum Chul Kwon, Ben Eysenbach, Janu Verma, Kenney Ng, Christopher deFilippi, Walter F. Stewart, and Adam Perer.

Presented by Jan Pilzer

Clustering

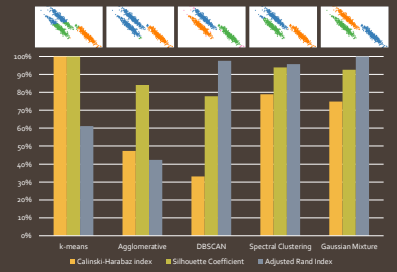
Unsupervised Clustering

Clustering Techniques



scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison

Clustering Metrics



Clustervision

The Joy of Clustering

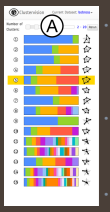
Design Goals

1. Compare clustering techniques and parameters
2. Compare clusters of a result
3. Compare data points within clusters
4. Understand the clustering
5. Steer clustering results

Overview of Clustervision

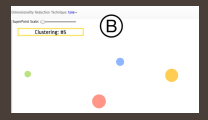
Dataset describing 403 paintings by the "Joy of Painting" artist Bob Ross.

Clustervision: Clustering Results



- Compute all possible combinations (58 results)
  - *k*-means, Spectral and Agglomerative Clustering
  - 19 parameter: k=2-20
- Analysed and ranked by clustering metrics
  - Calinski-Harabaz, Silhouette, Davies-Bouldin, S_{Dbw} and Gap Statistic
- Consistent colors for clusters

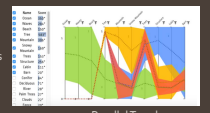
Clustervision: Projection



- Data points as circular elements in a two dimensional space, resembling a scatterplot
- Dimensionality reduction techniques to map into two dimensions
- Colors to represent cluster
- *Superpoints* to reduce visual clutter

Clustervision: Parallel Trends

- Rank features based on analysis of variance (ANOVA)
- Mean and 95% confidence intervals of features
- Option to sort and switch axes, and filter on features



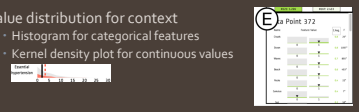
Parallel Trends

Clustervision: Cluster Detail

- Appears on selection of a cluster
- Summary of the clusters using statistics and prototypes
  - Cohesion: closeness of points in a cluster
  - Separation: distinctness of cluster to others
  - Silhouette: mean of silhouette scores
- Typical and atypical members
  - top 5 *inliers*: closest to center
  - top 5 *outliers*: farthest from center

Clustervision: Data Point

- Appears on selection of a point
- Details about actual values of features
- Value distribution for context
  - Histogram for categorical features
  - Kernel density plot for continuous values



Clustering Comparison

- Compare multiple clustering results
- Divide data items that are in different clusters in half
- Compare quality metrics directly

VAD Analysis

What: Data	Table with 67 categorical attributes
What: Derived	58 cluster assignments for each data item (one for each clustering)
Why: Tasks	Find correlation between attributes; Compare clustering results
How: Encode	Ranked List: Categorical hues on line marks and radar chart; Projection: Scatterplot; Parallel Trends: Parallel Coordinates using area marks for bundled lines; Cluster Detail: Column Chart; Data Point: Histogram and Kernel Density Plot
How: Facet	Multiform with linked highlighting and coloring; overview–detail with selection
Scale	403 paintings, 67 features, 58 clustering results

Case Study

Finding Clusters of Similar Patients

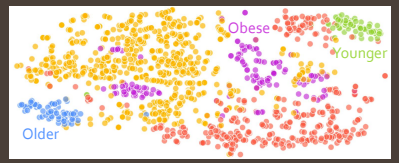
Previous Study Results

- 397 Patients diagnosed with HFpEF
- Hierarchical Clustering with k=1-8
- k=3 has highest score in Bayesian information criterion
- 3 archetypes of HFpEF
  - Younger patients, few comorbidities
  - Obese patients, diabetes
  - Older patients, chronic kidney disease

Clinically meaningful, but is there more?

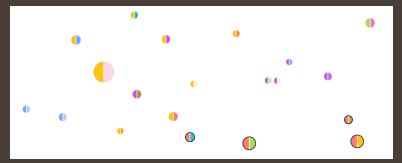
Study with Clustervision

- Data of HFpEF patients 2 years before diagnosis
- Hope for early treatments
- Results with k=3 do not map to previous study
- Result with k=5 has the 3 clusters of previous study



Study with Clustervision

- Two new clusters of younger and older patients
- Split red cluster by patients' medication
  - Teal: Calcium Channel Blockers and Loop Diuretics
  - Green: Thiazides and Thiazide-like Diuretics
  - Brown: ACE Inhibitors and Statins only
  - Gold: Statins, Ace Inhibitors, Beta Blockers, and Calcium-Channel Blockers



VAD Analysis

What: Data	Table with 23 attributes
What: Derived	Cluster assignments for each data item
Why: Tasks	Find correlation between attributes; Compare and evaluate clustering results
Scale	1474 patients, 23 features (comorbidities and medications)

Critique

Strengths

- Overview first, details-on-demand
  - Result List -> Scatterplot -> Cluster Info -> Point Info
- Consistent coloring for clusters
  - Between visualizations
  - Between results
- Good combination of existing idioms
- Parallel Trends as more readable version of parallel coordinates

Weaknesses

- Some features hidden
  - Cluster comparison on right click
  - Reordering and sorting not obvious (in screenshots)
- Implicit assumptions
  - Only show top 15 results (if significant difference)
  - Only show top 5 in- and outliers
- No radically new ideas

Resources

- Paper
  doi.org/10.1109/TVCG.2017.2745085
- Paper page with video
  bckwon.com/publication/clustervision
- Clustering algorithms and metrics in Python
  scikit-learn.org/stable/modules/clustering