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

<table>
<thead>
<tr>
<th>k-means</th>
<th>Agglomerative</th>
<th>DBSCAN</th>
<th>Spectral Clustering</th>
<th>Gaussian Mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="k-means" /></td>
<td><img src="image2" alt="Agglomerative" /></td>
<td><img src="image3" alt="DBSCAN" /></td>
<td><img src="image4" alt="Spectral Clustering" /></td>
<td><img src="image5" alt="Gaussian Mixture" /></td>
</tr>
</tbody>
</table>

[scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison](scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison)
Clustering Metrics

- k-means
- Agglomerative
- DBSCAN
- Spectral Clustering
- Gaussian Mixture

Metrics:
- Calinski-Harabaz index
- Silhouette Coefficient
- Adjusted Rand Index
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.
• 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 clusters
- *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
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

<table>
<thead>
<tr>
<th>What: Data</th>
<th>Table with 67 categorical attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>What: Derived</td>
<td>58 cluster assignments for each data item (one for each clustering)</td>
</tr>
<tr>
<td>Why: Tasks</td>
<td>Find correlation between attributes; Compare clustering results</td>
</tr>
<tr>
<td>How: Encode</td>
<td>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</td>
</tr>
<tr>
<td>How: Facet</td>
<td>Multiform with linked highlighting and coloring; overview–detail with selection</td>
</tr>
<tr>
<td>Scale</td>
<td>403 paintings, 67 features, 58 clustering results</td>
</tr>
</tbody>
</table>
Case Study
Finding Clusters of Similar Patients
• 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?
• 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

<table>
<thead>
<tr>
<th>What: Data</th>
<th>Table with 23 attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>What: Derived</td>
<td>Cluster assignments for each data item</td>
</tr>
<tr>
<td>Why: Tasks</td>
<td>Find correlation between attributes; Compare and evaluate clustering results</td>
</tr>
<tr>
<td>Scale</td>
<td>1474 patients, 23 features (comorbidities and medications)</td>
</tr>
</tbody>
</table>
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
• 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