Clustervision: Visual Supervision of Unsupervised Clustering

Kwon, B., Eysenbach, B., Verma, J., Ng, K., deFilippi, C., Stewart, W. F., and Perer, A.

Presented by Jan Pilzer

Clustering

Unsupervised Clustering

- Compare clustering techniques and parameters
- Compare clusters of a result
- Compare data points within clusters
- Understand the clustering
- Steer clustering results

Clustering Comparison

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

VAD Analysis

- Table with k categorical attributes
- Heatmap of k clusters and prototypes
- Influence of features on clusters

Why Tasks

- Finding clusters between attributes by highlighting
- Principal components analysis

What Enrich

- Barplot of clusters and prototypes
- Parallel coordinates plot
- Area marks for visualizing

How Enrich

- Balloon: k clusters and prototypes
- Parallel coordinates plot
- Area marks for visualizing

Cluster Data Point

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

Previous Study Results

- 397 Patients diagnosed with HFpEF
- Hierarchical Clustering with k=1-8
- k=3 has highest score in Bayesian information criterion (58 results)
- k=3 has lowest score in silhouette, Davies, and Gap Statistic
- Summary of clusters using statistics and prototypes
- Typical and atypical members
- Top 5 inliers: closest to center
- Top 5 outliers: farthest from center

Case Study

Finding Clusters of Similar Patients

- Younger
- Older

- Patients, diabetes
- Obese patients, chronic kidney disease
- Clinically meaningful, but is there more?

Overview of Clustervision

- Clustering: Visual Supervision of Unsupervised Clustering

Design Goals

- Clustervision: The Joy of Clustering

Clustervision

- 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: Projection

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

Clustervision: Parallel Trends

- Appears on selection of a cluster
- Summary of the clusters using statistics and prototypes
- Typical and atypical members
- Top 5 inliers: closest to center
- Top 5 outliers: farthest from center

Clustervision: Cluster Detail

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

Clustervision: Data Point

- 397 Patients diagnosed with HFpEF
- Hierarchical Clustering with k=1-8
- k=3 has highest score in Bayesian information criterion (58 results)
- k=3 has lowest score in silhouette, Davies, and Gap Statistic
- Summary of clusters using statistics and prototypes
- Typical and atypical members
- Top 5 inliers: closest to center
- Top 5 outliers: farthest from center

Clustering

Techniques

- Adjusted Rand Index
- Compare clustering techniques and parameters
- Spectral Clustering

Spectral Clustering

- Compare visual clutter
- Patients, chronic kidney disease

Older

- Table with 67 categorical attributes
- Cohesion: closeness of points in a cluster
- DBSCAN

Ranked List: Categorical hues on line

3 archetypes of clusters

Compare clusters of a result

Youth

- Silhouette Coefficient
- Gaussian Mixture
- Compare data points within clusters
- Top 5 parameter: k=2

Histogram for categorical features

Kernel density plot for continuous values

K=3 has highest score in Bayesian information criterion (58 results)

- Summary of clusters using statistics and prototypes
- Typical and atypical members
- Top 5 inliers: closest to center
- Top 5 outliers: farthest from center

Clinically meaningful, but is there more?
Study with Clustervision

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

- Two new clusters of younger and older patients
- Split red cluster by patients’ medication
  - Between visualizations
  - Between results
- Consistent coloring for clusters
- Parallel Trends as more readable version of parallel coordinates

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