Towards a Systemic Combination of Dimension Reduction and Clustering in Visual Analytics

John Wenskovitch, Ian Crandell, Naren Ramakrishnan, Leanna House, Scoland Leman, Chris North

Presented by Kaiyuan Li

Overall contributions
- An overview of combining dimension reduction and clustering techniques into a visualization system.
- A discussion of design decisions that must be addressed when creating a visualization system that combines two algorithms.

Overview of two algorithms
- Dimension reduction
  - To represent high-dimensional data in low-dimensional data in the meantime the properties and structure (outliers and clusters) of high-dimensional data can be preserved.
  - Advantage: scalability
  - Disadvantage: information loss
  - Linear and Nonlinear

Common used Dimension reduction algorithms
- Most common used: PCA

Overview of two algorithm-----clustering
- Clustering algorithm is usually for specific problems, no global optimal solutions.
  - Hierarchical --- divisive and agglomerative
  - Partitioning --- k-means

Distance function
- Distance function --- input of dimension reduction algorithm
- Measure the similarity for a pair of observations
- P-norms for more detailed: https://www.youtube.com/watch?v=EEca6f0eew
- P=1 Manhattan distance, P=2 Euclidean distance.
- Large dataset present preference difficulties, ASK-Graph view supports large dataset. (200000 nodes and 1600000 edges)

Tasks for Dimension reduction and Clustering
- Common goal: interaction and exploration in dataset
- Exploratory data analysis tasks —— gain insights
- Apply the weights to the dimension
- Overview of two algorithm-----
- Dimension reduction

Six combinations of Dimension Reduction and Clustering: pipeline examples
1> Independent Algorithm
- Execute indecendingly both algorithm without any influences
2> Dimension reduction preprocessing for Clustering
- Priorizing DRA first and some information of output pass to Clustering algorithm
3> Clustering preprocessing for dimension reduction
- Reverse process of previous pipeline
4> One algorithm implicitly includes the other
- Execute one of the algorithms, convert the output as the outputs of the other algorithm
5> Global and Local Algorithm Combination
- DRA take a global view and clustering algorithms take a local view, communicate with each other and converge to optimal layout
6> Iterative, alternative algorithm
- Work together in same overarching algorithm (K-means)

Visual representation properties
- Depending on algorithms (six pipelines)
  1. Represent cluster using convex hull, clearly show the different cluster
  2. May not produce optimal clustering on high-dimension
  3. Visibly separated clusters, the dimension reduction may not be optimal
  4. Inherent limitations depending on the algorithm applied
  5. Middle choice, overall layout effective, however not accurate as applying independent algorithms
  6. Both algorithms work simultaneously, near-optimal structure, however runtime are sacrificed

Interaction techniques
- PI (parametric Interaction)
- OLI (Observation-level Interaction)
- Surface-level

Design decision
- Node-link, space-filling, scatterplot, streamgraph
- Dealing with large dataset —— high risk: display (overdraw)
- Solution: abstract of observations into single glyph, filter the number of observations
Thank you, question ?