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

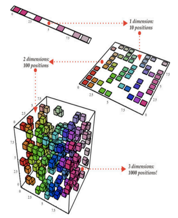
John Wenskovitch, Ian Crandell, Naren Ramakrishnan, Leanna House, Scoland Leman, Chris North

Presented by
Kaiyuan Li

Overall contributions

- an overview of combining dimension reduction and clustering techniques into a visualization system. (algorithm, task, visualization and interaction)
- A discussion of design decisions that must be addressed when creating a visualization system that combines two algorithm

Overview of two algorithm----- 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

	Selected Dimension Reduction Algorithms
Linear	Factor Analysis [43] Principal Component Analysis (PCA) [74] Probabilistic PCA (PPCA) [84] Projection Pursuit [40]
Both	Feature Selection [42] Independent Component Analysis (ICA) [49] Multidimensional Scaling (MDS) [85] Weighted MDS (WMDS) [18]
Nonlinear	Glimmer [50] Isomap [82] Latent Dirichlet Allocation (LDA) [11] t-Distributed Stochastic Neighbor Embedding (t-SNE) [65]

Most common used: PCA

Distance function

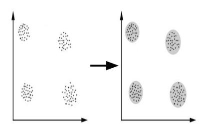
$$d_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_k |x_{i,k} - x_{j,k}|^p\right)^{1/p}.$$

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=_EEcjn0Uirw

P=1 Manhattan distance, p=2 Euclidean distance.

Large dataset present preference difficulties, ASK-Graph view supports large dataset. (200000 nodes and 16000000 edges)

Overview of two algorithm----- clustering



Clustering algorithm is usually for specific problems, no global optimal solutions

- Hierarchical ----divisive and agglomerative
- Partitioning----k-means

Preference difficulties in vary large dataset, preference improvement in small dataset (observations, dimensions)

P-F-R algorithm is designed for large dataset

Tasks for Dimension reduction and Clustering

	Dimension Reduction	Both	Clustering
See the Result	See distribution of observations	See relative positions of observations	Identify clusters of observations
Understand the Result	Measure distances between observations	Identify attribute values of observations	Label clusters Determine cluster structure
Affect the Result	Change distance metric Select different dimensions	Reposition observations in the full space Enhance an existing pattern in the projection	Change cluster membership of observations Create/remove clusters

Common goal: interaction and exploration in dataset

Exploratory data analysis tasks------gain insights
Apply the weights to the dimension

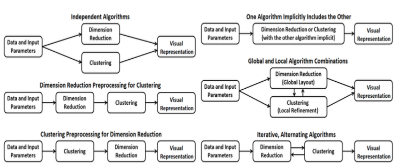
Coordinating two algorithms

Before selecting algorithms : what parameters should be learn and used

Distance function as the input is not all the same for algorithms, even if with same sets of weight.

It is impossibly to coordinate all pairs of dimension reduction algorithms and clustering algorithms

Six combinations of Dimension Reduction and Clustering: pipeline examples



1>Independent Algorithm
: execute indecently both algorithm without any influences

2>Dimension reduction preprocessing for Clustering
: processing DEA 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

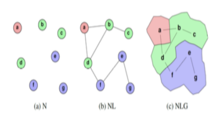
5>Global and local Algorithm Combination
: DRA take a global view and clustering algorithms take a local view, communicate with each other and converge to optimal layout

6>Iterative, alternative algorithm
:work together in same overarching algorithm (K-means)

Visual representation ----issues

Based on the algorithms, representing the results of computation

Node-link, space-filling, scatterplot, streamgraph



Dealing with large dataset ---------high risk i.. display (overdrawn)
Solution: abstract of observations into single glyph, filter the number of observations

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

	Dimension Reduction	Both	Clustering
PI	Rotate the projection	Modify the weight on a dimension Select a different distance function	Modify the max/min radius of a cluster Change the number of clusters sought
OLI	Reposition an observation external to clusters or within a single cluster	Reposition an observation into a different cluster	Change cluster membership Merge several clusters or split a cluster
Surface	Measure a distance between observations	Details-on-demand to obtain attribute values	Count the size of a cluster Annotate a cluster

Design decision

Section	Design Decision
2.1 & 2.2	In general, clustering places an emphasis on relationships within and between clusters. In contrast, dimension reduction emphasizes observation-to-observation relationships. Which of these tasks is the primary goal of the analyst?
3.2	What properties of the data is the visualization seeking to highlight? Which properties of the data are the system and analyst trying to discover? Should the primary goal of the visualization system be emphasizing observation relationships, clusters of observations, or both? Should the dimension reduction and clustering algorithms use the same distance function (if possible), or should each algorithm use an independent method of measuring similarity?
3.3	Which order and interaction of dimension reduction and clustering algorithms best models the task that the visualization system is addressing?
4.1	How can we encode distances and cluster membership information when both algorithms are present?
4.2	As the dimension reduction and clustering algorithms are competing in the same visualization, what features should be emphasized in the visualization to best address the problem?
5.2	Should interactions be designed independently for the dimension reduction and clustering algorithms, or should a given interaction affect both algorithms?

Thank you, question ?