Lecture 10: Attribute Reduction Methods
Information Visualization
CPSC 533C, Fall 2011

Tamara Munzner
UBC Computer Science
Wed, 12 October 2011

Required Readings

Chapter 8: Attribute Reduction Methods

Further Reading


Data Reduction

- how to reduce amount of stuff to draw?
- crosses view composition considerations
- item reduction
  - last time
  - rows of table
- attribute reduction
  - this time
  - columns of table
- methods for both
  - filtering, aggregation, ordering

Attribute Reduction Methods

- camera metaphors
  - slicing, cutting, projection
  - filtering, ordering, aggregation
- for attributes as opposed to items
- dimensionality reduction
  - uncovering hidden structure
- estimating true dimensionality
- generating synthetic dimensions
  - linear mappings
  - nonlinear mappings
  - displaying low-dimensional spaces
  - scatterplots, SPLOMS, landscapes

Slicing/Cutting: Spatial Data

- easy to understand: spatial data, 3D to 2D, axis aligned

Slicing: HyperSlice

- 4D function \[ \sum_{i=0}^{3} w_i (1 + \frac{|x - p_i|^2}{d}) \]
  - diagonals = standard graph

Slicing: High-Dimensional Functions

- HyperSlice: matrix of orthogonal 2D slices
  - each panel is display and control: drag to change slice
  - simple 3D example

Attribute Filtering

- filtering, but for attributes rather than items
- unfiltered vs filtered SPLOM

Attribute Ordering

- ordering, but for attributes rather than items
- Hierarchical Clustering Explorer

Dimensionality vs Attribute Reduction

- vocab use in field not consistent
- dimension/attribute
- attribute reduction: reduce set with filtering
  - includes orthogonal projection
  - dimensionality reduction: create smaller set of new dims
  - set size is smaller than original, new dims completely synthetic
  - clarification: includes dimensional aggregation
  - includes some projections (but not all)

Uncovering Hidden Structure

- measurements indirect not direct
- real-world sensor limitations
- measurements made in sprawling space
- documents, images
- DR only suitable if (almost) all information could be conveyed with fewer dimensions
- how do you know? need to estimate true dimensionality to check if different than original!

Estimating True Dimensionality

- error for low-dim projection vs high-dim original
- no single correct answer; many metrics proposed
- cumulative variance that is not accounted for
- strain: match variations in distance (vs actual distance values)
- stress: difference between interpoint distances in high and low dimensions

- stress(D, Δ) = \[ \sqrt{\frac{\sum_{i=1}^{n} (d_{ij} - Δ_{ij})^2}{D_{ij}^2}} \]
  - D: matrix of lowD distances
  - Δ: matrix of highD distances Δ_i^2

Showing Dimensionalities Estimates

- scree plots as simple way: error against i \( i \) dims
- original dataset: 294 dims
- estimate: almost all variance preserved with < 20 dims

Required Readings

- Attribute Reduction Methods
- Further Reading
- Dimensionality vs Attribute Reduction
- Slicing: HyperSlice
- Uncovering Hidden Structure
- Estimating True Dimensionality
- Showing Dimensionalities Estimates
Capturing Material Reflectance

- measurement: interaction of light with real materials (spheres)
- result: 104 high-res images of material
- each image 4M pixels

Goal: Image Synthesis

- step 1: create new renderings with CG objects that look like captured materials
- CG teapot looks just like real hematite

Need For Low-Dimensional Model

- how to do step 2 simulation of new materials?
- 104 materials * 4M pixels = 400 million dimensions
- model much too hi-dim to be useful

Dimensionality Reduction: Linear

- first try: PCA, linear DR technique
- result: error falls off sharply
- good results for step 1 around 45 dims
- step 2 problem: physically impossible intermediate points when simulating new materials
- specular highlights cannot have holes!

Finding Semantics for Synthetic Dimensions

- look for meaning in scatterplots
- each synthetic dimension named by people, not by algorithm
- points represent real-world images (spheres)
- people inspect images corresponding to points to decide if axis could have a meaningful name

Understanding Synthetic Dimensions

- Specular-Metallic
- Diffuseness-Glossiness

Nonlinear Dimensionality Reduction

- MDS, charting, Isomap, LLE, TSNE,...
- many techniques proposed
- problem: can handle curved rather than linear structure
- pros: can handle curved rather than linear structure
- new dimensions cannot be easily related to originals

Spring-Based MDS: Naive

- repeat for all points
- MDS: multidimensional scaling
- confusingly, large family of things all called MDS
- some linear, some nonlinear!
- classical: minimize strain
- early formulation equivalent to PCA (linear)
- spectral methods: approximate eigenvectors
- distance scaling: minimize stress
- nonlinear optimization
- force simulation (mass-spring)
Faster Spring Model: Stochastic
- compare distances only with a few points
- maintain small local neighborhood set
- each time pick some randoms, swap in if closer
- small constant: 6 locals, 3 randoms typical
- \( O(n) \) iteration, \( O(n^2) \) algorithm

Glimmer Algorithm
- multilevel to avoid local minima, designed to exploit GPU
- restriction to decimate
- relaxation as core computation
- relaxation to interpolate up to next level

Glimmer vs Stochastic Alone
- GPU version of stochastic as relaxation subsystem
- poor convergence properties if run alone
- only obvious when scalability allows thorough testing

Stochastic Termination
- how do you know when it’s done?
- no absolute threshold, depends on dataset
- interactive click to stop does not work for subsystem

GPUs
- characteristics
  - small set of localized texture accesses
  - output at predetermined locations
  - no variable length looping
  - avoid conditionals: all floating point units execute same instr at same time
  - mapping problems to GPU
  - arrays become textures
  - inner loops become fragment shader code
  - program execution becomes rendering

Finding/Verifying Clusters
- sparse document dataset: 28K dims, 28K points
- Glimmer (distance) vs PivotMDS (classical)
  - speed improvement so distance as fast as classical
  - major quality difference for sparse datasets

Showing DR Data
- scatterplot showing points
  - only works if true dimensionality is 2 (... or 3)
  - need to drill down to see what points represent
  - SPLOM
  - safe choice
  - landscapes
  - avoid! studies show worse than just using points

Reading For Next Time
Hierarchical Parallel Coordinates for Exploration of Large Datasets
Ying-Huey Fu, Matthew O. Ward, and Elke A. Rundensteiner,
IEEE Visualization ’99.
Metric-Based Network Exploration and Multiscale Scatterplot.

Reminders
- Project meetings due 10/19
- one week from today
- Office hours today after class (5-6)
- or schedule specific meeting time by email
- No class Oct 24/26