Interactive Visual Exploration of Multivariate Data Sets

Matthew O. Ward
Computer Science Department
Worcester Polytechnic Institute

work was supported under NSF Grant IIS-9732897

SC4DEVO-1, July 12-15, 2004
What is Multivariate Data?

• Each data point has N variables or observations
• Each observation can be:
  – nominal or ordinal
  – discrete or continuous
  – scalar, vector, or tensor
• May or may not have spatial, temporal, or other connectivity attribute
Sources of Multivariate Data

- Sensors (e.g., images, gauges)
- Simulations
- Census or other surveys
- Commerce (e.g., stock market)
- Communication systems
- Spreadsheets and databases
Purposes of Visualization

- Presentation of information/results
- Confirmation of hypotheses/analysis
- Exploration to develop model/hypothesis
Visual Tasks (from Keller&Keller)

- Identify
- Locate
- Distinguish
- Categorize
- Cluster
- Rank
- Compare
- Associate
- Correlate
- …
Methods for Visualizing Multivariate Data

• Dimensional Subsetting
• Dimensional Reorganization
• Dimensional Embedding
• Dimensional Reduction
Dimensional Subsetting

- Scatterplot matrix displays all pairwise plots
- Selection allows linkage between views
- Clusters, trends, and correlations readily discerned between pairs of dimensions
Dimensional Subsetting (2)

- Pixel-oriented techniques lay out a series of univariate displays
- Values are conveyed via color
- Records are ordered temporally, by value, or by a user query
Dimensional Reorganization

- Parallel Coordinates creates parallel, rather than orthogonal, dimensions.
- Data point corresponds to polyline across axes
- Clusters, trends, and anomalies discernable as groupings or outliers, based on intercepts and slopes
Dimensional Reorganization (2)

- Glyphs map data dimensions to graphical attributes
- Size, color, shape, and orientation are commonly used
- Similarities/differences in features give insights into relations
Dimensional Embedding

- Dimensional stacking divides data space into bins
- Each N-D bin has a unique 2-D screen bin
- Screen space recursively divided based on bin count for each dimension
- Clusters and trends manifested as repeated patterns
Dimensional Reduction

- Map N-D locations to M-D display space while best preserving N-D relations
- Approaches include MDS, PCA, and Kohonen Self Organizing Maps
- Relationships conveyed by position, links, color, shape, size, etc.
The Role of Interaction

- User needs to interact with display, examine interesting patterns or anomalies, validate hypotheses
- Selection allows isolation of subset of data for highlighting, deleting, focused analysis
- Navigation allows alternate views, drill-down for details
- Direct (clicking on displayed items) vs. indirect (range sliders, text queries)
- Screen space (2-D), data space (N-D), structure space (spatio-temporal, grids, hierarchies)
Problems with Large Data Sets

- Most techniques are effective with small to moderate sized data sets
- Large sets (> 50K records) are increasingly common
- When traditional visualizations used, occlusion and clutter make interpretation difficult
Examples of Scale Problem
Common Approaches to the Problem of Scale

- Sampling
- Filtering
- Aggregation and Summarization
- Dimensionality Reduction (e.g., PCA, MDS)
- Binning
- Multiresolution Methods***
Multiple Resolutions in Visual EDA

• For each target (number of records, dimensions, distinct nominal values)
  – Apply hierarchical clustering algorithm
  – Identify representative value for each non-terminal cluster
  – Compute cluster descriptors to convey contents
  – Visualize representative values using traditional tools, augmented with descriptors
  – Provide interactive tools to navigate, modify, and filter the hierarchical structure
Visualizing Large Numbers of Records: Mean-Band Method

- User specifies focus region in data space and level of detail for focused/unfocused areas
- Mean value for each cluster displayed in color based on its location in hierarchy
- Opacity bands around data points show population and extent of clusters
Hierarchical Parallel Coordinates

- Bands show cluster extents in each dimension
- Opacity conveys cluster population
- Color similarity indicates proximity in hierarchy
Hierarchical Scatterplots

- Clusters displayed as rectangles, showing extents in 2 dimensions
- Color/opacity consistently used for relational and population info
Navigating Hierarchies

- Drill-down, roll-up operations for more or less detail
- Need selection operation to identify subtrees for:
  - Exploration
  - Manipulation
  - Pruning
- Can be user-driven, data-driven, structure-driven
Structure-Based Brushing

- Enhancement to screen-based and data-based methods
- Specify focus, extents, and level of detail
- Intuitive - wedge of tree and depth of interest
- Implemented by labeling/numbering terminals and propagating ranges to parents
Structure-Based Brush

- White contour links terminal nodes
- Red wedge is extents selection
- Color curve is depth specification
- Color bar maps location in tree to unique color
- Direct and indirect manipulation of brush
Visualizing Large Numbers of Dimensions: VHDR

- User specifies multiple foci in hierarchical dimension space and level of detail for each
- Visualizations convey representative dimensions and local (for each data record) and global (for all dimensions in cluster) degree of dissimilarity in cluster
Manipulating Hierarchical Structures via InterRing

Dimension hierarchy composed of 4 dimensions
InterRing: Hierarchy Modification

- **Goal:** change hierarchy manually
- **Interaction:** drag and drop
- **Traceability:** color preserving

WPI Center for Research in Exploratory Data and Information Analysis

SC4DEVO-1, July 12-15, 2004
Selecting Clusters for Viewing

- **Goal**: select clusters from hierarchy
- **Manual brushing**: select each cluster by mouse click
- **Structure-based brushing**: select multiple clusters at one time according to clustering parameter
A Sample Session
Load a Data Set
Cluster Dimensions
Examine Subsets of Dimensions
Find Redundant, Uninformative Dimensions
Select Diverse Dimensions
Display, Alter Dimensions if Desired
Highlight Subsets, Find Patterns
Change Views and Iterate
Zoom In on Dimensions
Summary

• Hierarchical/multiresolution techniques one solution to problem of scale
• Can be inter-record, inter-dimension, or intra-dimension
• For each, need:
  – Method(s) to generate hierarchies
  – Method(s) to summarize hierarchies
  – Method(s) to visually convey hierarchies
  – Methods to interact (navigation, selection)
• All need to be easy to understand and control
Current and Future Work

• Automated view refinement to reduce clutter and enhance visual structure
• Integration of quality attributes for data values, dimensions, and records – quality management, visualization, and interaction
• Performance and scalability – how much data is needed in order to make decisions
• Merging analytic and visual data mining
More …

- XmdvTool has been in the public domain since 1994.
- XmdvTool website:
  - http://davis.wpi.edu/~xmdv/
  - Contains:
    - source code
    - build environments for Windows, Linux, and Unix
    - Windows and Linux executable
    - Documentation, paper reprints, and case studies
    - Data sets
Questions?