Multivariate Visual Representations 1

CS 4460 – Intro. to Information Visualization
Sep. 11, 2014
John Stasko

Agenda

• General representation techniques for multivariate (>3) variables per data case
  – But not lots of variables yet...
Quick Quiz

• What type of dataset has three variables per case?
• What is a scatterplot matrix?

How Many Variables?

• Data sets of dimensions 1, 2, 3 are common
• Number of variables per class
  – 1 - Univariate data
  – 2 - Bivariate data
  – 3 - Trivariate data
  – >3 - Hypervariate data Focus Today
Earlier

- We examined a number of tried-and-true techniques/visualizations for presenting multivariate (typically $\leq 3$) data sets
  - Hinted at how to go above 3 dimensions

Representations

Some standard ways for low-d data

Tukey box plot
Hypervariate Data

- How about 4 to 20 or so variables (for instance)?
  - Lower-dimensional hypervariate data
  - Many data sets fall into this category
  - Often modeled as tables or tabular data

More Dimensions

- Fundamentally, we have 2 geometric (position) display dimensions
- For data sets with >2 variables, we must project data down to 2D
- Come up with visual mapping that locates each dimension into 2D plane

- Computer graphics: 3D->2D projections
Wait a Second

- A spreadsheet already does that
  - Each variable is positioned into a column
  - Data cases in rows
  - This is a projection (mapping)

- What about some other techniques?
  - Already seen a couple

Multiple Views

Give each variable its own display

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>
Scatterplot Matrix

Represent each possible pair of variables in their own 2-D scatterplot

Chernoff Faces

Encode different variables’ values in characteristics of human face
Examples

Cute applets:
http://www.cs.uchicago.edu/~wiseman/chernoff/

Table Lens

- Spreadsheet is certainly one hypervariate data presentation
- Idea: Make the text more visual and symbolic
- Just leverage basic bar chart idea

Rao & Card
CHI’94
Visual Mapping

<table>
<thead>
<tr>
<th>Name</th>
<th>Salary</th>
<th>Quota</th>
<th>Variance to quota% of quota</th>
<th>Forecast</th>
<th>Actual bookings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alicia</td>
<td>200,000</td>
<td>130,000</td>
<td>-10,000</td>
<td>200,000</td>
<td>100,000</td>
</tr>
<tr>
<td>Brian</td>
<td>150,000</td>
<td>80,000</td>
<td>157</td>
<td>200,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Catherine</td>
<td>200,000</td>
<td>150,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>600,000</td>
</tr>
<tr>
<td>David</td>
<td>180,000</td>
<td>120,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>900,000</td>
</tr>
<tr>
<td>Emily</td>
<td>200,000</td>
<td>140,000</td>
<td>-60,000</td>
<td>200,000</td>
<td>200,000</td>
</tr>
<tr>
<td>Frank</td>
<td>160,000</td>
<td>100,000</td>
<td>-60,000</td>
<td>200,000</td>
<td>300,000</td>
</tr>
<tr>
<td>Gary</td>
<td>140,000</td>
<td>80,000</td>
<td>157</td>
<td>200,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Helen</td>
<td>200,000</td>
<td>150,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>600,000</td>
</tr>
<tr>
<td>Ian</td>
<td>180,000</td>
<td>120,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>900,000</td>
</tr>
<tr>
<td>Jennifer</td>
<td>200,000</td>
<td>140,000</td>
<td>-60,000</td>
<td>200,000</td>
<td>200,000</td>
</tr>
<tr>
<td>Karen</td>
<td>160,000</td>
<td>100,000</td>
<td>-60,000</td>
<td>200,000</td>
<td>300,000</td>
</tr>
<tr>
<td>Lisa</td>
<td>140,000</td>
<td>80,000</td>
<td>157</td>
<td>200,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Matthew</td>
<td>200,000</td>
<td>150,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>600,000</td>
</tr>
<tr>
<td>Nicole</td>
<td>180,000</td>
<td>120,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>900,000</td>
</tr>
<tr>
<td>Olivia</td>
<td>200,000</td>
<td>140,000</td>
<td>-60,000</td>
<td>200,000</td>
<td>200,000</td>
</tr>
<tr>
<td>Peter</td>
<td>160,000</td>
<td>100,000</td>
<td>-60,000</td>
<td>200,000</td>
<td>300,000</td>
</tr>
<tr>
<td>Rachel</td>
<td>140,000</td>
<td>80,000</td>
<td>157</td>
<td>200,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Sarah</td>
<td>200,000</td>
<td>150,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>600,000</td>
</tr>
<tr>
<td>Tom</td>
<td>180,000</td>
<td>120,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>900,000</td>
</tr>
<tr>
<td>Uma</td>
<td>200,000</td>
<td>140,000</td>
<td>-60,000</td>
<td>200,000</td>
<td>200,000</td>
</tr>
<tr>
<td>Victor</td>
<td>160,000</td>
<td>100,000</td>
<td>-60,000</td>
<td>200,000</td>
<td>300,000</td>
</tr>
<tr>
<td>William</td>
<td>140,000</td>
<td>80,000</td>
<td>157</td>
<td>200,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Xander</td>
<td>200,000</td>
<td>150,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>600,000</td>
</tr>
<tr>
<td>Yianni</td>
<td>180,000</td>
<td>120,000</td>
<td>-20,000</td>
<td>200,000</td>
<td>900,000</td>
</tr>
<tr>
<td>Zara</td>
<td>200,000</td>
<td>140,000</td>
<td>-60,000</td>
<td>200,000</td>
<td>200,000</td>
</tr>
<tr>
<td>Total</td>
<td>1,000,000</td>
<td>600,000</td>
<td>157</td>
<td>200,000</td>
<td>500,000</td>
</tr>
</tbody>
</table>

Change quantitative values to bars

Tricky Part

What do you do for nominal data?
Instantiation

Details

Focus on item(s) while showing the context
See It


Video

FOCUS

• Feature-Oriented Catalog User Interface
• Leverages spreadsheet metaphor again
• Items in columns, attributes in rows
• Uses bars and other representations for attribute values

Spenke, Beilken, & Berlage
UIST ’96
Characteristics

- Can sort on any attribute (row)
- Focus on an attribute value (show only cases having that value) by double-clicking on it
- Can type in queries on different attributes to limit what is presented too
Categorical data?

- How about multivariate categorical data?

- Students
  - Gender: Female, male
  - Eye color: Brown, blue, green, hazel
  - Hair color: Black, red, brown, blonde, gray
  - Home country: USA, China, Italy, India, ...
Mosaic Plot

Women

Men

Fall 2014  CS 4460  25

Fall 2014  CS 4460  26
Mosaic Plot

Women

Men

Brown
Hazel
Green
Blue

Mosaic Plot

Black  Red  Brown  Blond

Women

Men

Brown
Hazel
Green
Blue
Attribute Explorer

- General hypervariate data representation combined with flexible interaction

Characteristics

- Multiple histogram views, one per attribute (like trellis)
- Each data case represented by a square
- Square is positioned relative to that case’s value on that attribute
- Selecting case in one view lights it up in others
- Query sliders for narrowing
- Use shading to indicate level of query match (darkest for full match)
Features

- Attribute histogram
- All objects on all attribute scales

- Interaction with attributes limits

Features

- Inter-relations between attributes – brushing
Features

- Color-encoded sensitivity

Attribute Explorer

Video

Summary

- Summary
  - Attribute histogram
  - Attribute relationship
  - Sensitivity information
  - Especially useful in “zero-hits” situations or when you are not familiar with the data at all

- Limitations
  - Limits on the number of attributes

MultiNav

- Each different attribute is placed in a different row
- Sort the values of each row
  - Thus, a particular item is not just in one column
- Want to support browsing
Alternate UI

- Can slide the values in a row horizontally
- A particular data case then can be lined up in one column, but the rows are pushed unequally left and right
Attributes as Sliding Rods

Instantiation

Video

https://www.youtube.com/watch?v=GEBx-XTrGps
Limitations

- Number of cases (horizontal space)
- Nominal & textual attributes don’t work quite as well

Parallel Coordinates

- What are they?
  - Explain...
**Parallel Coordinates**

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>D2</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>D3</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
Parallel Coordinates

V1  V2  V3  V4  V5  D2
2    7    6    3    4

Parallel Coordinates

V1  V2  V3  V4  V5  D3
9    8    1    4    2
Parallel Coordinates

Encode variables along a horizontal row

Vertical line specifies different values that variable can take

Data point represented as a polyline

Parallel Coords Example

Basic

Grayscale

Color
Issue

- Different variables can have values taking on quite different ranges
- Must normalize all down (e.g., 0->1)

Application

- System that uses parallel coordinates for information analysis and discovery
- Interactive tool
  - Can focus on certain data items
  - Color

Taken from:
A. Inselberg, "Multidimensional Detective"
InfoVis '97, 1997.
Discuss

• What was their domain?
• What was their problem?
• What were their data sets?

The Problem

• VLSI chip manufacture
• Want high quality chips (high speed) and a high yield batch (% of useful chips)
• Able to track defects
• Hypothesis: No defects gives desired chip types
• 473 batches of data
The Data

- 16 variables
  - X1 - yield
  - X2 - quality
  - X3-X12 - # defects (inverted)
  - X13-X16 - physical parameters

Parallel Coordinate Display

Yikes!
But not that bad

Figure 1: The full dataset consisting of 473 batches
**Top Yield & Quality**

- Defects
- Split

Figure 2: The batches high in Yield, X1, and Quality, X2.

Have some defects

**Minimal Defects**

- Not the highest yields and quality

Figure 3: The batches with zero in 9 out of the ten defect types.
Best Yields

Appears that some defects are necessary to produce the best chips

Non-intuitive!

XmdvTool

Toolsuite created by Matthew Ward of WPI

Includes parallel coordinate views
ParVis System

http://www.mediavirus.org/parvis/

Challenges

Out5d dataset (5 dimensions, 16384 data items)

courtesy of J. Yang
Dimensional Reordering

Which dimensions are most like each other?

Same dimensions ordered according to similarity

Yang et al
InfoVis '03

Dimensional Reordering

Can you reduce clutter and highlight other interesting features in data by changing order of dimensions?

Peng et al
InfoVis '04
Reducing Density

Jerding and Stasko, '95, '98
Wegman & Luo, '96

Artero et al, 04

Johansson et al, '05

Improved Interaction

• How do we let the user select items of interest?

• Obvious notion of clicking on one of the polylines, but how about something more than that
Attribute Ratios

- Angular Brushing
  - Select subsets which exhibit a correlation along 2 axes by specifying angle of interest

  Hauser, Ledermann, & Doleisch
  InfoVis '02

(earlier demo)

Range Focus

- Smooth Brushing
  - Specify a region of interest along one axis
Different Kinds of Data

- How about categorical data?
  - Can parallel coordinates handle that well?
Parallel Sets

- Visualization method adopting parallel coordinates layout but uses frequency-based representation
- Visual metaphor
  - Layout similar to parallel coordinates
  - Continuous axes replaced with boxes
- Interaction
  - User-driven: User can create new classifications

Kosara, Bendix, & Hauser
TVCG '05

Representation

Color used for different categories

Those values flow into the other variables
Example

Titanic passengers data set

<table>
<thead>
<tr>
<th>Class</th>
<th>Sex</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td></td>
<td>145 44.6%</td>
<td>180 55.4%</td>
<td>325 14.8%</td>
</tr>
<tr>
<td></td>
<td>30.8% 6.6%</td>
<td>10.4% 8.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>second</td>
<td></td>
<td>106 37.2%</td>
<td>179 62.8%</td>
<td>285 12.9%</td>
</tr>
<tr>
<td></td>
<td>22.6% 4.8%</td>
<td>10.4% 8.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>third</td>
<td></td>
<td>196 27.8%</td>
<td>510 72.2%</td>
<td>706 32.1%</td>
</tr>
<tr>
<td></td>
<td>41.7% 8.9%</td>
<td>29.5% 23.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>crew</td>
<td></td>
<td>23 2.6%</td>
<td>862 97.4%</td>
<td>885 40.2%</td>
</tr>
<tr>
<td></td>
<td>4.9% 1.1%</td>
<td>49.8% 39.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>470 21.4%</td>
<td>1731 78.6%</td>
<td>2201 100%</td>
</tr>
</tbody>
</table>

Titanic Data Set
Interactions

Fig. 7. Basic interaction elements in Parallel Sets: reordering categories (a, b) helps to generate a more meaningful layout; grouping categories (c, d) enables a hierarchical analysis/exploration; excluding categories from the visualization (e, f) allows for interactive filtering; and category highlighting (g, h) enables the selective investigation of high-dimensional relations.

Video

Fall 2014
Star Plots

Space out the n variables at equal angles around a circle

Each “spoke” encodes a variable’s value

Alternative Rep.

Data point is now a “shape”

Star Plot examples

http://seamoney.ed.asu.edu/~behrens/asu/reports/compre/compl.html
Star Coordinates

- Same ideas as star plot
- Rather than represent point as polyline, just accumulate values along a vector parallel to particular axis
- Data case then becomes a point


Demo
**Star Coordinates**

- Data cases with similar values will lead to clusters of points
- (What’s the problem though?)
- Multi-dimensional scaling or projection down to 2D

**Generalizing the Principles**

- General & flexible framework for axis-based visualizations
  - Scatterplots, par coords, etc.
- User can position, orient, and stretch axes
- Axes can be linked
Parallel Coordinates

- Technique
  - Strengths?
  - Weaknesses?
Project

- Teams & Topics due Tuesday
  - Bring 2 copies

- More topic ideas

HW 2

- Design table & graph
- Due Tuesday
  - Bring 2 hardcopies
Upcoming

- Multivariate Visual Representations 2
  - Reading: Munzner chapter 12

- D3 tutorial
  - Reading
    Interactive Data Visualizations for the Web, chapters 3 and 5

Application

Application