3. Multidimensional Information Visualization I

Concepts for visualizing univariate to hypervariate data

Dr. Thorsten Büring, 8. November 2007, Vorlesung Wintersemester 2007/08
Outline

- Reference model and data terminology
- Visualizing data with < 4 variables
- Visualizing multivariable data
  - Geometric transformation
  - Glyphs
  - Pixel-based
  - Dimensional Stacking
  - Downscaling of dimensions
- Case studies: support for exploring multidimensional data
  - Rank-by-feature
  - Value & relation display
  - Dust & magnet
- Clutter reduction techniques

Topics of next lecture: Multidimensional Information Visualization II
Information Visualization

- The use of computer-supported, interactive, visual representations of abstract data to amplify cognition (Card et al. 1999)
- How to construct interactive visual representations?
- Reference Model for Visualization

**Reference Model for Visualization**

Data → Visual Form

- **Data**: Raw Data, Data Tables, Visual Structures
- **Visual Form**: Views, Visual Mappings, View Transformations

**Human Interaction**

- **Raw Data**: idiosyncratic formats
- **Data Tables**: relations (cases by variables) + meta-data
- **Visual Structures**: spatial substrates + marks + graphical properties
- **Views**: graphical parameters (position, scaling, clipping, …)

Card et al. 1999
Data Table

- Cases (observations)
- Variables (aka attributes)
- Example car data set
  - 406 cases
  - 8 variables for each case
- Metadata
  - Descriptive information about the data
  - Units, e.g. lbs., mph, inches
  - Constraints, e.g. if var1is ‘41’, then var7 can only be ‘11’ or ‘3’
  - Data types

<table>
<thead>
<tr>
<th></th>
<th>Variableₓ</th>
<th>Variableᵧ</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caseᵢ</td>
<td>Valueᵢₓ</td>
<td>Valueᵢᵧ</td>
<td>...</td>
</tr>
<tr>
<td>Caseⱼ</td>
<td>Valueⱼₓ</td>
<td>Valueⱼᵧ</td>
<td>...</td>
</tr>
<tr>
<td>CaseⱿ</td>
<td>ValueⱿₓ</td>
<td>ValueⱿᵧ</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>mpg</td>
<td>cylinders</td>
<td>engine displ.</td>
<td>horsepower</td>
</tr>
<tr>
<td>Chevrolet C. M.</td>
<td>18</td>
<td>8</td>
<td>307</td>
</tr>
<tr>
<td>Datsun PL510</td>
<td>27</td>
<td>4</td>
<td>97</td>
</tr>
<tr>
<td>Audi 100 LS</td>
<td>24</td>
<td>4</td>
<td>107</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Dimensionality of Data

- On how many variables was a data case measured?
  - 1 variable – Univariate
  - 2 variables – Bivariate
  - 3 variables – Trivariate
  - Above 3 variables – Hypervariate aka multivariate aka multivariable data
- Visualizations that encode multivariable data are called multidimensional visualizations
- Visualizing multivariable data is one of the most challenging tasks in Information Visualization
Data Types

- **Nominal (categorical)**
  - Unordered set
  - Operators: =, ≠
  - Example: car origin (Europe, USA, Asia)

- **Ordinal**
  - Possess a natural order
  - Operators: <, >
  - Example: ratings, school grades

- **Quantitative**
  - Allow for arithmetic operations
  - Operators: *, /, +, -
  - Example: acceleration in seconds

- Also subtypes exist: e.g. quantitative geographic (geographic coordinates), quantitative time
Data Transformation

- Transformation of raw data into data tables can involve loss or gain of information
  - Classing: quantitative to ordinal data by dividing values into ranges, e.g. acceleration into <slow, medium, fast>
  - Nominal to ordinal data by sorting the values lexicographically
  - Derived values e.g. calculating statistical summaries (mean, median...)
  - Derived structures (e.g. sorting cases and / or variables)
  - Sampling (determining a representative subset of the data set)
  - Aggregation of data (e.g. determining frequencies)
- Deal with errors, missing values and duplicates
Objectives of Visual Structures

- Various mappings possible

- Quality factors of mapping
  - Expressiveness - all and only the data in the data table are represented in the structure
  - Increased effectiveness compared to another mapping
    - Faster to interpret
    - Can convey more distinctions
    - Leads to fewer errors in interpretation
    - See previous lecture on perception!

Card et al. 1999
Outline

- Reference model and data terminology
- Visualizing data with < 4 variables
- Visualizing multivariable data
  - Geometric transformation
  - Glyphs
  - Pixel-based
  - Dimensional Stacking
  - Downscaling of dimensions
- Case studies: support for exploring multidimensional data
  - Rank-by-feature
  - Value & relation display
  - Dust & magnet
- Clutter reduction techniques

Topics of next lecture: Multidimensional Information Visualization II
Univariate Data

- **Horsepower**

  - **Plot**
    - **Low**
    - **Middle 50%**
    - **High**
    - **Mean**
    - **Boxplot**
    - **Histogram**

- **Line graph - not very reasonable in this case**

- **Bargram**
Frequency Distribution Analysis

Images from Field & Hole 2003
Interactive Bargrams

- Demo
Bivariate Data

- Most common for displaying bivariate data is the scatterplot.
- Each spatial dimension is assigned a (usually quantitative) axis variable.
- Cases are mapped to a spatial position according to the data values for the axes.
- Users can easily identify global trends, local trade-offs, outliers ...
- Potential problems?
Scatterplot Analysis

- No relationship
- Strong linear (positive correlation)
- Homoscedastic
- Exact linear (positive correlation)
- Strong linear (negative correlation)
- Heteroscedastic

Scatterplot Analysis

- Quadratic relationship
- Exponential relationship
- Sinusoidal relationship (damped)
- Outlier

Time-Based Bivariate Data

- Plot of time series
  - X-axis represents time
  - Y-axis a function of time
- Closing prices of 1,430 individual stocks across 52 weeks of time

TimeSearcher, Hochheiser & Shneiderman 2004
Time Map

  - X-axis: month
  - Y-axis: years and weekdays (Sunday to Saturday)
  - 4 categories of ozone concentration mapped to distinct colors
- Reveals seasonal patterns
  - Ozone levels are much higher in summer months
  - High ozone days have steadily decreased
- How could this visualization be improved?

Image taken from Mintz et al. 1997
Geographic Bivariate Data

- Size of each territory shows relative proportion of the world population living there
- Potential problem with this visualization?

Image taken from worldmapper.org
Distorted Map with Comparison Map

Image taken from Spence 2007
Trivariate Data

- Tempting: map each variable to each dimension of a 3D scatterplot
- Occlusion of points with different positions
- Problem with static representation?
Scatterplot Matrix

- Matrix of all pairwise scatterplot views of the data
- Easy to understand by using familiar and powerful scatterplot representation
- Can serve as a good starting point for data exploration
- Increased demand for display space
- Increased cognitive load caused by redundant data

Cleveland 1993
Trivariate Data

- 2D scatterplot with additional encoding
- In this case color and shape
- Shows relationship between three variables
- For color / shape coding: assumes categorical variable or classing of quantitative variable (loss of information)
Outline

- Reference model and data terminology
- Visualizing data with < 4 variables
- Visualizing multivariable data
  - Geometric transformation
  - Glyphs
  - Pixel-based
  - Dimensional Stacking
  - Downscaling of dimensions
- Case studies: support for exploring multidimensional data
  - Rank-by-feature
  - Value & relation display
  - Dust & magnet
- Clutter reduction techniques

Topics of next lecture: Multidimensional Information Visualization II
Geometric Transformations

- Idea: present projections of the multidimensional data to find interesting correlations

- Most common techniques
  - Scatterplot matrix
  - Prosection matrix
  - Parallel coordinates plot
Scatterplot Matrix

- Scatterplot matrix can be scaled to > 3 variables
- Number of scatterplots increases rapidly
- n variables means n x n plots
- Diagonal maps the same variable twice
- Each pair is plotted twice, once on each side the diagonal
- Allows convenient sequential browsing of one variable compared to all other variables
Prosection Matrix

- Scatterplot matrix with **interactive linking and brushing** (Tweedie & Spence 1996)
- Projection of a section of parameter space
- User select multivariable ranges, which are colored differently
Parallel Coordinate Plot

- One vertical axis for each variable
- Every case is represented by a graph
- Graph intersects each of the vertical axis at the point corresponding to the attribute value of the case
- Popular visualization technique
- Complexity (number of axes) is directly proportional to the number of attributes (comp. scatterplot matrix)
- All attributes receive uniform treatment
- Demo
- Potential problems of this visualization?

Inselberg 1997
Parallel Coordinate Plot

- Bendix et al. 2005: Parallel Sets
- Parallel coordinates for categorical data
- Substitute individual data points by a frequency-based representation
- Any problems with this visualization?

<table>
<thead>
<tr>
<th>Sex</th>
<th>Class</th>
<th>1st</th>
<th>2nd</th>
<th>3d</th>
<th>crew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (s)</td>
<td>1st</td>
<td>141</td>
<td>93</td>
<td>90</td>
<td>3</td>
</tr>
<tr>
<td>Female (d)</td>
<td>2nd</td>
<td>4</td>
<td>13</td>
<td>106</td>
<td>20</td>
</tr>
<tr>
<td>Male (s)</td>
<td>3rd</td>
<td>62</td>
<td>25</td>
<td>98</td>
<td>670</td>
</tr>
<tr>
<td>Male (d)</td>
<td>crew</td>
<td>118</td>
<td>154</td>
<td>422</td>
<td>192</td>
</tr>
</tbody>
</table>
3D Parallel Coordinates

Parallel 2D planes instead of vertical axes

Parallel Coordinate Plot

Try it out

- XmdvTool  http://davis.wpi.edu/%7Exmdv/index.html
- Parvis  http://home.subnet.at/flo/mv/parvis/index.html
Geometric Transformations

- Advantages
  - Users’ familiarity with scatterplots (scatterplot matrix)
  - 2D patterns can easily be identified

- Disadvantages
  - Rather limited scalability
    - Number of cases (Parallel Coordinate Plot)
    - Number of dimensions (scatterplot matrix)
  - Overplotting and overlap
  - Labeling (Parallel Coordinates)
Glyph-Based Visualizations

Glyph-based techniques
- Star glyph
- Chernoff faces
- Stick-figure
- Shape coding
- Color icons

Glyph: small-sized visual symbol
- Variables are encoded as properties of glyph
- Each case is represented by a single glyph
Star glyphs

- Coekin 1996
- Radial axes with equal angles
- Each axis represents a variable
- Each spoke encodes a variable’s value
- May also be overlaid for better comparison

![Star glyphs diagram](http://www.itl.nist.gov/div898/handbook/eda/section3/starplot.htm)
Chernoff Faces

- Chernoff 1973
- Humans are sensitive to a wide range of facial characteristics (e.g. eye size, length of a nose, etc.)
- 18 characteristics to encode data by stylized faces
- Positive evaluation results (Spence & Parr 1991)
- Some facial features seem to be able to carry more information than others (Morris et al. 1999; De Soete 1986)
Stick-Figure Icons

- Pickett & Grinstein 1998
- Each case is represented by a stick figure
- Two attributes are mapped to XY position of the glyph
- Remaining dimensions are mapped to the angle and / or length of the 4 limbs
- When icons are densely packed a texture appears
- Texture pattern reveals characteristics of the data space
- Different members of stick-figure family for conveying different types of data structures
Stick-Figure Icons

- Stick-figure example
- Census data showing age (y), income (x), education, salary, language, marital status etc.
- Gender is encoded by two stick-figure families

Grinstein et al. 1989
3D Stick-Figures

Keim 2000
Shape Coding

- Beddow 1990
- Each case is drawn as a glyph containing a rectangular grid
- Each grid cell represents one attribute
- Attribute value is encoded with gray scales
- Glyphs are positioned in a line, columns or encoded dimensions
- Highly compressed visualization without clutter and overlap (compare to stick-figures)
- Identification of promising patterns
Shape Coding

- Attribute values encoded by white, grey, black
- 13 Variables gained from magnetosphere and solar wind data
- Includes one time variable (hour/day), which has been mapped to x/y

Figure 1: Day by Hour: Thirteen Parameters of Magnetosphere and Solar Wind Data
Color icons

- Levkowitz 1991, Keim & Kriegel 1994
- Shape coding with a focus on colors
- Arrangement is query-dependent (e.g. spiral: most relevant glyph is centered)
- What about compressing the visualization even more by using 1-pixel representations?
- Problem: users need at least 2x2 pixel per data value + pixels for borders to distinguish between the elements of the visualization
- This is different to pixel-based techniques, which will be discussed in the next lecture

Keim & Kriegel 1994
Keim 1994
Glyph-Based Visualizations

### Advantages
- Provide holistic overview of the information space
- Exploit the human powerful ability of perceiving (texture) patterns and human face characteristics (Chernoff)
- Direct metaphor of Chernoff-face-like icons (e.g. houses) may prove to be intuitive for novice users

### Disadvantages
- Glyphs must be learned
- Only suitable for small to medium data sets
- Stick figures give a rather broad overview and may be difficult to interpret
- Mappings may introduce biases in interpretation (e.g. the head shape of a Chernoff-face may be easier to perceive and compare than length of nose)