Visualization Techniques in Data Mining

Prof. Pier Luca Lanzi
Laurea in Ingegneria Informatica
Politecnico di Milano
Polo di Milano Leonardo
Outline

• Goals of visualization
• Advantages
• Methodologies
• Techniques
• User interaction
• Problems
Goals of Data Visualization

• Today there is the need to manage a huge amount of data, and computer systems help us in this task

• Visual Data Mining help to deal with this flood of information, integrating the human in the data analysis process

• Visual Data Mining allows the user to gain insight into the data, drawing conclusions and directly interacting with the data
Advantages of visualization techniques

The main advantages of the application of Visual data mining techniques are:

• Visual data exploration can easily deal with very large, highly non homogeneous and noisy amount of data

• Visual data exploration requires no understanding of complex mathematical or statistical algorithms

• Visualization techniques provide a qualitative overview useful for further quantitative analysis
Approach methodologies

**Presentation:**
- starting point: facts to be presented are fixed a priori
- result: high-quality visualization of the data presenting the facts

**Confirmative Analysis:**
- starting point: hypotheses about the data
- result: visualization of the data allowing confirmation or rejection of the hypotheses

**Explorative Analysis:**
- starting point: data without hypotheses
- result: visualization of the data, which can provide hypotheses about data distribution
Visualization techniques

- **Geometric techniques**: scatterplots matrices, Hyperslice, parallel coordinates
- **Pixel-oriented techniques**: simple line-by-line, spiral and circle segments
- **Hierarchical techniques**: Treemap, cone trees
- **Graph-based techniques**: 2D and 3D graph
- **Distortion techniques**: hyperbolic tree, fisheye view, perspective wall
- **User interaction**: brushing, linking, dynamic projections and rotations, dynamic queries
Geometric techniques

Basic idea:
• Visualization of geometric transformations and projections of the data

Methods:
• Scatterplot matrices
• Hyperslice
• Parallel coordinates
Scatterplot matrices

• A scatterplot matrix is composed of scatter plots of all possible pairs of variables in a dataset.

• Assuming a N-dimensional dataset, there are \((N^2-N)/2\) pairs of two dimension plots.
Hyperslice

• HyperSlice is an extension of the scatterplot matrix

• They represent a multi-dimensional function as a matrix of orthogonal two-dimensional slices
Parallel Coordinates

• The axes are defined as parallel vertical lines separated

• A point in Cartesian coordinates correspond to a polyline in parallel coordinates

• Able to visualize data that may be occluded in Cartesian coordinates
Pixel-oriented techniques

Basic idea:
• The basic idea of pixel-oriented techniques is to map each data value to a colored pixel
• Each attribute value is represented by a pixel with a color tone proportional to a relevance factor in a separate window

Methods:
• Simple Arrangement Line-by-Line
• Spiral and Circle Segments Techniques
Pixel-oriented techniques

- Simple arrangement line-by-line
Pixel-oriented techniques

• Spiral

• Circle segments
Hierarchical techniques

Basic idea:
Visualization of the data using a hierarchical partitioning into two- or three-dimensional subspaces

Methods:
• Treemap
• Cone trees
Treemaps

• Visualization of hierarchical collections of quantitative data as files on a hard drive, financial analysis, bioinformatics, etc..

• Divide a limited screen space display area into a sequence of rectangles whose areas correspond to an attribute of data set

http://www.smartmoney.com/marketmap/
Cone trees

3-dimensional extension of the more familiar 2-D hierarchical tree structures, to a more intuitive navigation and display of information.
Graph-based visualization

- Graphs (edges + nodes) with labels and attributes
- Used where emphasis is on data relationship (databases, telecom)
- Coordinates not always meaningful
- Useful for discovering patterns
Graph-based visualization

• Color and thickness code values

• Asymmetric relations:
Graph-based visualization

- E-mail (SeeNet)
Graph-based visualization

• 3D graphs:
  – more room for objects
  – different points of view

• Example (hypertexts – Narcissus):
Focus vs. context

• Too much data in too small screens
• Solutions:
  – dual views (detailed + global)
  – distorted view (e.g. fisheye view)
Distortion

- Hyperbolic tree

- Fisheye view

- Perspective wall
User interaction

• Brushing: selecting points or regions
• Linking: more views work together
User interaction

• Dynamic projections and rotations
  – Interactively and continuously moving through subspaces

• Dynamic queries
  – Visual interface (button and sliders)
  – Incremental behavior (undo)
Problems

• Missing attributes
  – Ignore
  – Fill blanks with:
    • a predefined constant
    • a value extracted according to the inferred distribution
  – Assess the effect of interpolated values
Problems

• Large data sets
  – Typical screens have one million pixels
  – Subsampling
  – Voxel/pixel bins
  – Jittering

• Large number of attributes
  – Principal component analysis
  – Factor analysis
  – Etc.
Conclusions

• Human and computer skills can be integrated with visual data mining

• Visualization may be useful for:
  – understanding what is happening
  – searching novel patterns

• User interaction is paramount in these
References (I)

References (II)

- J. Lamping, R. Rao, P. Pirolli. “A focus+context technique based on hyperbolic geometry for visualizing large hierarchies”. CHI ’95, pp. 401-408