Part II

Scatter plots
Outline

4 Introduction

5 Feature number reduction
   - Principles
   - Neighborhood structure preservation

6 Overlapping reduction
   - Rendering and interaction
   - Clustering
“Old fashioned” methods from exploratory data analysis:

- bar chart and histogram
- boxplot
- scatter plot
- run chart
- etc.

They are still useful!
Don’t use pie chart (never ever! no kidding!)

Don’t use pie chart (never ever! no kidding!)

Scatter plot

The standard tool

- one point per object
- two dominant numerical characteristics per object
- a few additional characteristics:
  - a nominal variable (hue or shape coded)
  - a numerical variable (lightness or shape coded)
  - a label
- doesn’t scale:
  - low dimension only
  - very sensitive to overlapping
Abalone dataset
Abalone dataset

![Abalone (UCI) diagram](image)

- **Shell weight** vs **Diameter**
- Legend:
  - **Sex**:
    - Female
    - Infant
    - Male

The diagram shows a scatter plot of the Abalone dataset from the UCI Machine Learning Repository. The x-axis represents the diameter, while the y-axis represents the shell weight. The points are color-coded by sex:
- Female (△)
- Infant (○)
- Male (▼)

The distribution of points suggests a correlation between shell weight and diameter, with males generally having larger shell weights at any given diameter compared to females and infants.
Abalone dataset

Abalone (UCI)

Sex
- Female
- Infant
- Male

Diameter

Shell weight
How to improve scalability?

Two scalability problems:

1. **Dimension (a.k.a. variable or characteristics or feature):**
   - Reduce the number of dimension:
     - user choice
     - automated: selection and extraction
   - Display more than 2/3 dimensions at once: visual layout

2. **Object:**
   - Reduce the number of objects: clustering and quantization
   - Reduce the size of an object
   - Constrain the display to forbid (or reduce) overlapping
Outline

4 Introduction

5 Feature number reduction
   - Principles
   - Neighborhood structure preservation

6 Overlapping reduction
   - Rendering and interaction
   - Clustering
Automated scalability

Reducing the number of variables (Machine Learning):

- A major research topic
- Feature selection: keeping a subset of the original variables
- Feature extraction: producing new variables in small quantity
- Supervised: variables should “explain” a target variable
- Or Unsupervised: variables optimize some quality criterion

Some methods:

- Principal Component Analysis
- Linear Discriminant Analysis
- Sammon’s Mapping
- Variable Clustering
- etc.

see Prof. Lee’s lecture
Difficulties

What should be optimized?

- Goal dependent:
  - outlier detection $\Rightarrow$ maximize distances between outliers and central objects
  - visual clustering $\Rightarrow$ respect the neighborhood relationships
  - rule finding $\Rightarrow$ keep original variables
  - etc.

- Machine learning algorithms optimize abstract quantities (e.g., reconstruction error)

- Links between ML optimality and visual usefulness is unclear (at best!)
Italian Wines (PCA 99.98 %)
Linear projections

Italian Wines (LDA)
Linear projections
Linear projections

![Spheres Diagram](image-url)
Linear projections

Spheres (PCA 69 %)
Linear projections

Spheres (LDA)

LD1

LD2
Neighborhood structure preservation

To enable visual cluster analysis on projected dataset, the neighborhood structure of the data must be preserved by the projection.

- Quality measures (Venna & Kaski, 2001 → 2007):
  - Trustworthiness: neighbors on the screen are real neighbors
  - Continuity: real neighbors are neighbors on the screen
  - Precision $\approx$ Trustworthiness
  - Recall $\approx$ Continuity

- Optimization methods:
  - Stochastic Neighbor Embedding (Hinton & Roweis, 2002)
  - Neighbor Retrieval Visualizer (Venna & Kaski, 2006)
Neighbor Preservation

Original space to projection space

6 neighbors
Neighbor Preservation

Original space to projection space

Correct projection
Neighbor Preservation

Original space to projection space

Trustworthiness violation
Neighbor Preservation

Original space to projection space

Continuity violation
Trustworthiness and Precision

Can you trust neighbors in the projection space?

- $O_k(x_i)$: $k$-nn of $x_i$ in the original space
- $P_k(x_i)$: $k$-nn of $x_i$ in the projection space
- $U_k(x_i) = P_k(x_i) \setminus O_k(x_i)$

**Precision**

- maximal precision: $P_k(x_i) \subset O_k(x_i)$
- mean on $i$ of $1 - \frac{\#U_k(x_i)}{\#P_k(x_i)}$

**Trustworthiness**

- rank preservation: $r^O(x_j, x_i)$ rank of $x_j$ as a neighbor of $x_i$ in the original space
- $M_1(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{x_j \in U_k(x_i)} \left(r^O(x_j, x_i) - k\right)$
Contribution to precision: 0.5
Contribution to trustworthiness: 9
Continuity and Recall

Do you miss neighbors in the projection space?

- \( O_k(x_i) \): \( k \)-nn of \( x_i \) in the original space
- \( P_k(x_i) \): \( k \)-nn of \( x_i \) in the projection space
- \( V_k(x_i) = O_k(x_i) \setminus P_k(x_i) \)

Recall

- maximal recall: \( O_k(x_i) \subset P_k(x_i) \)
- mean on \( i \) of \( 1 - \frac{\#V_k(x_i)}{\#O_k(x_i)} \)

Continuity

- rank preservation: \( r^P(x_j, x_i) \) rank of \( x_j \) as a neighbor of \( x_i \) in the projection space

\[
M_2(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{x_j \in V_k(x_i)} (r^P(x_j, x_i) - k)
\]
Contribution to recall: 0.5
Contribution to continuity: 8
Stochastic Neighbor Embedding (SNE)

Approximating the “neighborhood distribution” (Hinton & Roweis, 2002)

- Original space: \( p_{ij} = \frac{\exp(-d_{ij}^2)}{\sum_{k \neq j} \exp(-d_{ik}^2)} \) (\( d_{ij} \) dissimilarity between \( x_i \) and \( x_j \))
- \( x_i \) projected to \( y_i \)
- Projection space: \( q_{ij} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq j} \exp(-\|y_i - y_k\|^2)} \)
- Minimize the Kullback-Leibler divergence between \( p_i \) and \( q_i \), i.e.,

\[
C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}
\]

- Corresponds to a smoothed version of the recall (Venna & Kaski, 2006)
An example

Italian Wines (PCA 99.98 %)

PC1

PC2

F. Rossi (INRIA)

Infovis & ML

27-28/02/2007

68 / 172
An example

Italian Wines (SNE)

V1

V2

-0.5 0.0 0.5 1.0 1.5
-0.5 0.0 0.5 1.0 1.5
No miracle
No miracle
No miracle

Spheres (SNE)

V1

V2

−0.5 0.0 0.5 1.0 1.5
−0.5 0.0 0.5 1.0 1.5
Optimizing smoothed versions of precision and recall (Venna & Kaski, 2006)

- inspired by stochastic neighbor embedding (SNE)
- minimization of

\[ \lambda \sum_{i} \sum_{j} p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - \lambda) \sum_{i} \sum_{j} q_{ij} \log \frac{q_{ij}}{p_{ij}} \]

- \( \lambda = 1 \) (SNE) \( \Rightarrow \) recall
- \( \lambda = 0 \) \( \Rightarrow \) precision
- \( \lambda \) is user chosen (can be chosen to optimize a quality measure)
- slow \( (O(N^3) \) per iteration for \( N \) objects, as SNE)
Neighbor Retrieval Visualizer (No class info!)

Spheres (NeRV, $\lambda=0$)
Neighbor Retrieval Visualizer (No class info!)

Spheres (NeRV, $\lambda=0.2$)
Neighbor Retrieval Visualizer (No class info!)

Spheres (NeRV, $\lambda = 0.5$)
Neighbor Retrieval Visualizer (No class info!)
Outline

4 Introduction

5 Feature number reduction
   • Principles
   • Neighborhood structure preservation

6 Overlapping reduction
   • Rendering and interaction
   • Clustering
Reducing the overlapping problem

An open research topic.

- **Rendering:**
  - Transparency
  - stereo vision
  - overlap counting (via openGL!)

- **Interactivity:**
  - sub-sampling
  - zooming and panning
  - excentric labeling
  - magic lenses
  - distortion

- **AI:**
  - clustering
  - force field approach
Transparency

- Dots are not fully opaque
- Overlapping produces:
  - Enhanced saturation for identical original colors
  - Blended colors for distinct original colors
- Similar idea: overlap counting (left) of the scatter plot (right)

http://www.cs.umd.edu/hcil/millionvis/
Transparency

![Standard Scatter Plot]

- Transparency in visualization.
- Information visualization and machine learning.

---

F. Rossi (INRIA)  Infovis & ML  27-28/02/2007  75 / 172
Excentric labeling

Applet Viewer: Test7.class

Radius: 25
Circle
Rect
X Stable
Y Stable
Alpha Stable

Name: Hans-Dieter Koch

# random items:

Applet démarré.
Excentric labeling
Excentric labeling
Excentric labeling
Excentric labeling
Clustering

Reduction of the number of objects:
- Very general solution (not limited to scatter plot)
- Clustering $\Rightarrow$ Cluster prototype $\Rightarrow$ Prototype visualization

Interactivity:
- Number of clusters:
  - deterministic clustering
  - least disturbance (e.g., hierarchical)
- Cluster extension (objects or summary)
Hierarchical clustering

Abalone (UCI)

Sex
- Female (217.8)
- Infant (223.7)
- Male (218.3)

Diameter vs. Shell weight plot.
Hierarchical clustering

Abalone (UCI)

Sex
- Female (100.5)
- Infant (103.2)
- Male (101.9)
Hierarchical clustering

Abalone (UCI)

Sex
- Female (20.1)
- Infant (20)
- Male (20.1)

Diameter
Shell weight

F. Rossi (INRIA)
Infovis & ML
27-28/02/2007
Hierarchical clustering

Abalone (UCI)

Sex
- Female (10.1)
- Infant (10)
- Male (10.1)
Hierarchical clustering

Abalone (UCI)

Sex
- Red: Female (5)
- Green: Infant (5)
- Blue: Male (5)

Shell weight vs. Diameter

- X-axis: Diameter
- Y-axis: Shell weight

The scatter plot shows the relationship between shell weight and diameter for different sexes of abalones. The data points are color-coded by sex, with female abalones in red, infant abalones in green, and male abalones in blue. The plot illustrates the clustering of data points by sex, with males generally having a higher shell weight for a given diameter compared to females and infants.
Hierarchical clustering

Abalone (UCI)

Sex
- Female (2)
- Infant (2)
- Male (2)

Shell weight vs. Diameter
Abalone (UCI)

Sex
- Female (217.8)
- Infant (223.7)
- Male (218.3)
Cluster content

Abalone (UCI)

Diameter vs. Shell weight

Sex

- Female (217.8)
- Infant (223.7)
- Male (218.3)
Abalone (UCI)

Sex
- Female (217.8)
- Infant (223.7)
- Male (218.3)
Hierarchical clustering

Abalone (UCI)

Diameter

Shell weight

Rings

Sex

- Female (217.8)
- Infant (223.7)
- Male (218.3)
Hierarchical clustering

Abalone (UCI)

- Sex
  - Female (100.5)
  - Infant (103.2)
  - Male (101.9)

- Rings
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7
  - 8
  - 9
  - 10
  - 11
  - 12
  - 13
  - 14
  - 15
  - 16
  - 17
  - 18
  - 19
  - 20
  - 21
  - 22
  - 23
  - 24
  - 25
  - 26
  - 27
  - 29

- Diameter
- Shell weight
- Rings

F. Rossi (INRIA)  Infosvis & ML  27-28/02/2007
Hierarchical clustering

Abalone (UCI)

Shell weight

Sex
- △ Female (20.1)
- ○ Infant (20)
- ▽ Male (20.1)

Rings
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29

Diameter
Hierarchical clustering

Abalone (UCI)

- Rings
- Sex
  - Female (10.1)
  - Infant (10)
  - Male (10.1)

Shell weight vs. Diameter

F. Rossi (INRIA)
Infovis & ML
27-28/02/2007
Hierarchical clustering

Abalone (UCI)

Diameter vs. Shell weight

Rings
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29

Sex
Female (5) Infant (5) Male (5)

F. Rossi (INRIA)
Hierarchical clustering

Abalone (UCI)

Rings

Sex

- Female (2)
- Infant (2)
- Male (2)
Hierarchical clustering and Transparency

Abalone (UCI)

- **Diameter**
- **Shell weight**
- **Rings**
- **Sex**
  - Female (217.8)
  - Infant (223.7)
  - Male (218.3)
Hierarchical clustering and Transparency

Abalone (UCI)

Sex
- Female (100.5)
- Infant (103.2)
- Male (101.9)
Hierarchical clustering and Transparency

Abalone (UCI)

Diameter
Shell weight
Rings
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
29

Sex
- Female (20.1)
- Infant (20)
- Male (20.1)
Hierarchical clustering and Transparency

Abalone (UCI)

Diameter
Shell weight
Rings
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
29

Sex
Female (5)
Infant (5)
Male (5)

F. Rossi (INRIA)
Infovis & ML
27-28/02/2007
Hierarchical clustering and Transparency

Abalone (UCI)

Diameter
Shell weight
Rings

Sex
Female (2)
Infant (2)
Male (2)

F. Rossi (INRIA)
Infovis & ML
27-28/02/2007
Hierarchical clustering and Transparency

Abalone (UCI)

- **Rings**: Colors indicate the number of rings, ranging from 1 to 29.
- **Sex**:
  - Female: △
  - Infant: ○
  - Male: ▽

- **Shell weight** and **Diameter** are plotted on the y-axis and x-axis, respectively.

F. Rossi (INRIA)  Infovis & ML  27-28/02/2007  81 / 172