Summary Visual Analytics

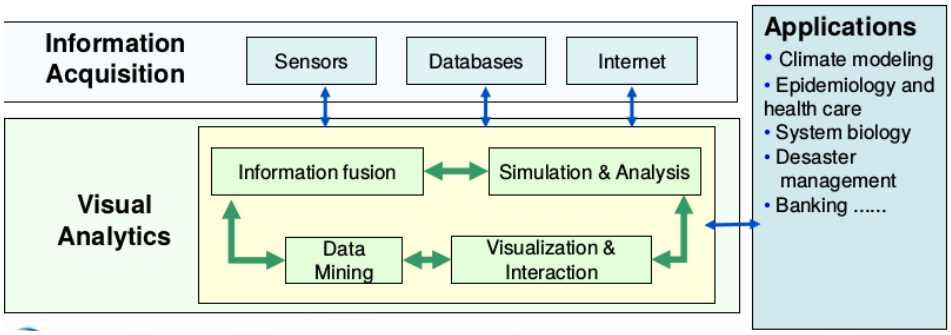
John

April 20, 2017

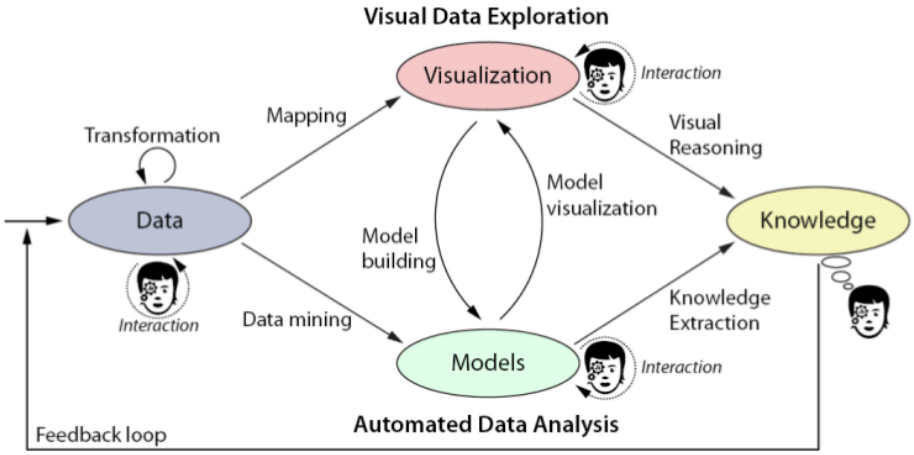
Contents

1 Introduction

- Visual Analytics is the science of analytical reasoning facilitated by interactive visual interfaces.
- An integrated combination of data analytics and interactive visual exploration.
- aims at supporting analysts
- Information is no more the bottleneck. Instead, analytical capabilities are essential.
- insights about, trends, patterns and relations as well as new hypothesis



- The scaleability challenge: information, variables, display, human
- Components
  - Analytical reasoning How to maximise human capacity to perceive, understand, and reason about complex and dynamic data and situations?
  - Visual representations and interaction techniques How to augment cognitive reasoning with perceptual reasoning through visual representations and interaction?
  - Data representations and transformations How to transform data into a representation that is appropriate to the analytical task and effectively conveys the important content?
  - Production, presentation, and dissemination How to convey analytical results in meaningful ways to various audiences? (Also to justify the efforts of visual analytics)



- Filtering operates on data values and includes
  - Data cleansing
  - Analytical computations, e.g. statistics

- reproducible results
- Major applications
  - Business and finance
  - Emergency management
  - Security
  - Sport analytics
  - Astronomy
  - Network analytics
  - Climate and weather research

2 Clustering

Clustering is part of an exploratory data analysis where users have little a priori information and serves to define groups automatically (unsupervised learning).

What is clustering good for?

- Learn the structure of data, e.g. define subgroups for customer segmentation (business analytics)
- A preprocess for selective visualization (show only certain clusters) or for focus-context visualization (show cluster representatives as overview and all instances as detailed view)
- A preprocess for classification, e.g. as input for a decision tree search

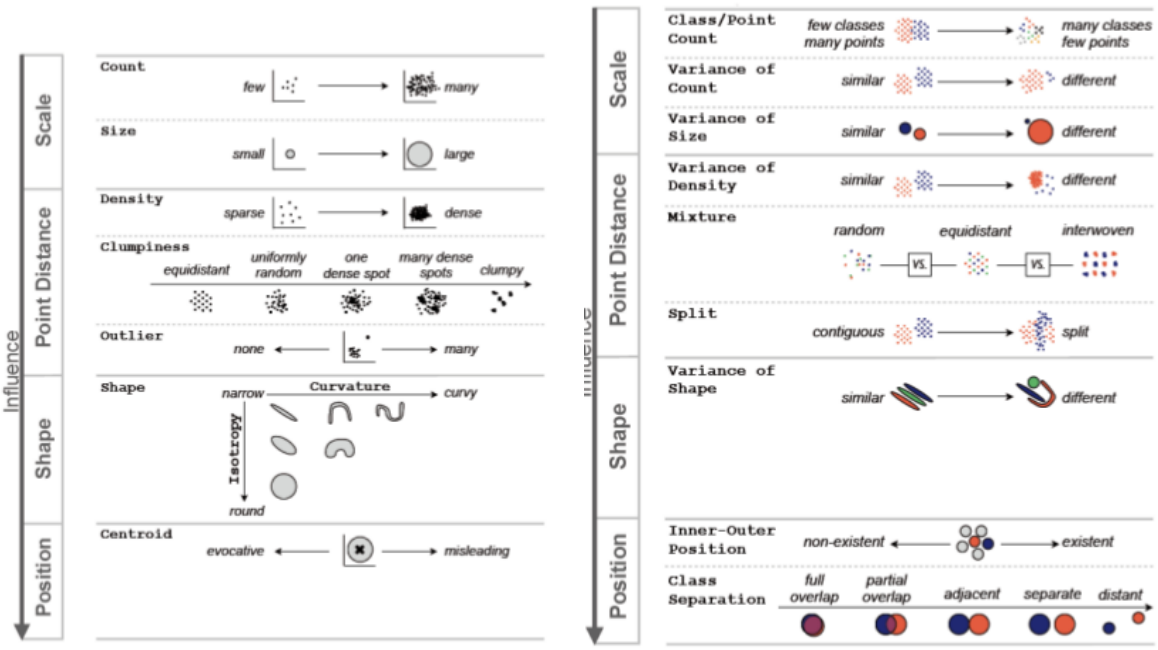
According to the model, clustering is performed

- In a hierarchic or non-hierarchic manner
- In a fuzzy or binary manner (hard)
- In a deterministic or non-deterministic manner
- Using various distance measures

Outliers (not belonging to any cluster) are possible with some approaches.
It is easier to interactively merge clusters (right image) than to separate an erroneously connected cluster.

2.1 Clustering Methods

- K-Means
- Optics
- Density-Based
- Agglomerative Hierarchical
- Clustering Ensembles



→ Fuzzy/Hard clustering

Clustering may be based on

- A distance model
- A density model
- A hierarchy model where clusters are assumed to exist at different levels

k-means:

- Partioning of a dataset in k groups
- follows a centroid model (a cluster is defined by its center)
- random initialization
- convex clusters
- not robust
- k must be known
- not deterministic

Fuzzy c-means

- With fuzzy clustering objects on the boundaries between cluster centers are not forced to fully belong to one of them, but are assigned membership degrees between 0 and 1 indicating partial memberships

Density base clustering

- do not require an a priori number of expected clusters and generate clusters of arbitrary shapes.

- Optics and DB-SCAN
- Basic concept: density-connectivity → two objects are density connected , if there is a chain of dense objects that connect the objects

Hierarchical clustering

- generates a hierarchy of clustering results
- Elements are connected and assigned to a cluster, if their distance is below a threshold
- If the threshold is increased, low level clusters get connected to higher level clusters
- Thus, a hierarchy arises bottom up (agglomerative)
- Essential parameters of hierarchical clustering are:
  - Distance function, e.g. Euclidean, Manhattan, Mahalanobis distance
  - Linkage criterion
  - The method is not suitable for large element numbers

The results of clustering may be improved by with a priori knowledge: Must-link constraints, Cannot-link constraints.
Clustering in general is an unsupervised method; clustering with constraints is supervised

Temporal Clustering

3 Subspace Clustering

4 Clustering Validation

5 BiClusters

6 Scatter plots

7 Dimension Reduction

8 Association Rule Mining

9 Decision Trees

10 Regression Based Analysis

11 Temporal Event Sequences

12 Visual Analytics Healthcare

13 Interactive cooperative Visual Analytics