Clustering

K. Gibert

Knowledge Engineering and Machine Learning group at

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, (IDEAI-UPC) Barcelona

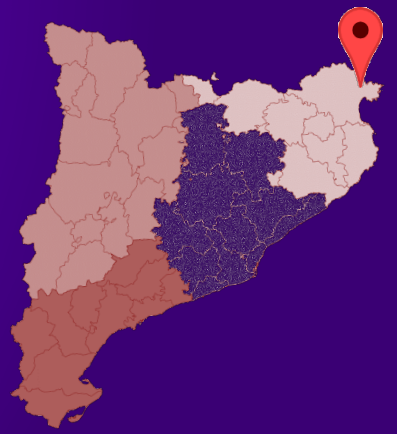
Dean of Illustrious Professional College of Informatics Engineering of Catalonia

President and founder of donesCOEINF

https://www.eio.upc.edu/en/homepages/karina,

karina.gibert@upc.edu, karina.gibert, @karinagibertk

Figuerenca



Planeta , Abril, 2024

Finding distinguishable groups with homogeneous individuals

basic brain activity

First systematic trial: LINNEO (s. XVII)

- Formal solutions
  - Statistics
  - Artificial Intelligence

Different background

Clustering

- Impossibility theorem of Kleinberg *[Kleinberg 2003]*
  - Given a clustering function f assigning classes to objects
  - *Scale-Invariance* (classes are maintained by distance scaling)
  - *Richness* The algorithm can produce all P(I) by changing parameters
  - *Consistency*: clusters are invariant by Γ-transformations
  - cannot be hold simultaneously
  -
- Trade-offs inherent to clustering problem
- Many relaxations provide different problems

Kleinberg, J. M. (2003). An impossibility theorem for clustering. In Advances in neural information processing systems (pp. 463-470).

Clustering

- Decomposition of variability

Huygens Theorem $I = B + W$

$$I = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$(x_i - \bar{x}) = (x_i - \bar{x}_c) + (\bar{x}_c - \bar{x})$$

$$2(x_i - \bar{x}_c)(\bar{x}_c - \bar{x}) = 0$$

$$W = \sum_{i=1}^{n} (x_i - \bar{x}_c)^2$$

$$B = \sum_{c=1}^{C} (\bar{x}_c - \bar{x})^2$$

Clustering

- Optimization problem

$$Max\ \{min\ d_{Within}\ (x_i - \bar{x}_c) + \ \lambda min\ d_{Between}\ (\bar{x}_c - \bar{x}\ \ )\}$$

- Searching space dimension

$$\sum_{k=1}^{n}\left(\frac{1}{k!}\sum_{i=0}^{k}(-1)^{(k-i)}\binom{k}{i}i^n\right)$$

n=100, k=3, Partitions= 10^{47}

n=25, k=5, Partitions= 10^{19}

n=100, k=5, Partitions= 10^{68}

Heuristic criteria
Required
1000 algorithms (2015)

Clustering
Statistical principles

- Algebraic fundamentals
  - *Only numerical data matrices*
  - Sokal and Sneath 1956 Numerical Taxonomy

Distance required

- Partitioning methods (linear complexity)
  - Number of classes IS AN INPUT
  - K-means *[McQueen67]*, dynamic clouds (nuées dynamiques, Diday)

- Hierarchical methods (quadratic complexity)
  - Number of classes IS AN OUTPUT
  - Ascendents or descendents (for very large n)

Curse of dimensionality

- Bad performance if large number of variables (compensation effect)
- *A huge "normal " group and many outlier groups (trivial knowledge)*

Clustering
Artificial Intelligence principles

- Logic and information theory fundamentals

  Often qualitative data matrices

  - Conceptual clustering (Michalski & Stepp 1983)
    - COBWEB (Fisher 1987)
    - ITERATE (Biswas 1998)

  - Fuzzzy clustering
    - Fuzzy C-Means (Bezdek 1981)

Clustering
Model based approaches

- Probabilistic clustering:
  - *Assume known initial distributions for classes*
  - *EM-algorithm: Two step*
    - ✓ Expectation step: Compute the expetected class of objects (use conditional distributions and posterior probabilities)
    - ✓ Maximization step: Update distributional class parameters to maximize the current class assignments
    - ✓ (use likelihood functiion, update distributional parameers)
    - ✓ Repeat till no improvement
  - ✓ Generative topographic mapping [Bishop 1995]

Distributional assumptions

Convergence not guaranteed

Optimal not guaranteed

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86-97.
Madhulatha, T. S. (2012). An overview on clustering methods. arXiv preprint arXiv:1205.1117. Jain AK, Dubes RC (1998) Algorithms for clustering data. Prentice Hall Inc.
Michalski, R. S., & Stepp, R. E. (1983). Learning from observation: Conceptual clustering. In Machine learning (pp. 331-363). Springer Berlin Heidelberg.
Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. Machine learning, 2(2), 139-172.
Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203.

Clustering
Model based approaches

- Density Estimation based
  - *Search areas with higher concentration of observations over data cloud*
  - *Assume density homogeneity and some parameters*

Estornells (Starling) https://www.youtube.com/watch?v=ZJBVHptmcO4

Cigonyes (Grus/Stork) https://www.youtube.com/watch?v=V3501Bdi4Oo

https://www.youtube.com/watch?v=AYXktkfMnSI
https://www.youtube.com/watch?v=oqMS5fn0LbM
https://www.youtube.com/watch?v=7Ddnw5Ln-aw

Nature and density based patterns

Nature and density based patterns

Clustering
Other approaches

- Neural-networks based
  - SOM *[Kohonen, 1998]*

- Collaborative methods
  - Multiview *[Bickel 2004] [Sevilla-Villanueva 2017]*
  - Probabilistic collaborative clustering *[Forestier 2010]*
    (la matrice des probabilities)

Kohonen, T. (1998). The self-organizing map. Neurocomputing, 21(1-3), 1-6.

Sevilla-Villanueva, B., Gibert, K., & Sànchez-Marrè, M. (2017). A methodology to discover and understand complex patterns: Interpreted Integrative Multiview Clustering (I2MC). Pattern Recognition Letters, 93, 85-94.

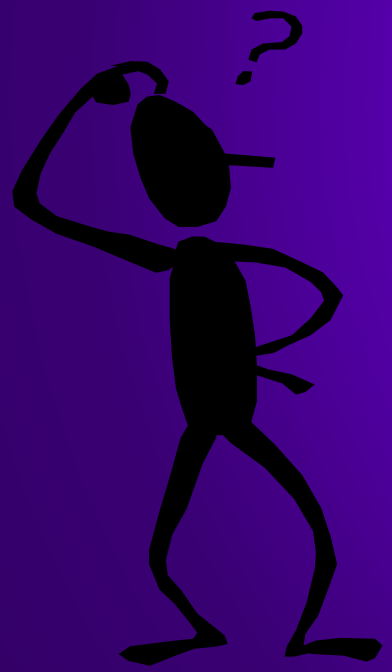
S. Bickel , T. Scheffer , Multi-view clustering, in: ICDM, 4, 2004, pp. 19–26 .

G. Forestier, P. Gançarski, C. Wemmert. Collaborative clustering with background knowledge, Data & Knowledge Engineering, 2010.

Clustering

- Clustering based on rules *[Gibert 1996]*:
  - *Sea sponges [LNStats1994] [Mathware 1997]*
  - *Stellar populations [CyS 1998]*
  - *Thyroid dysfunctions [JAMSDA 1999]*
  - *Characteristic situations in wastewater treatment plants [AIComm2001, 2005]*
  - *Reaction time after electroshock therapy [LNCS2002] [MedicinskaInformatika 2003]*
  - *Response to antidepressants treatment in patients with schizophrenia [ENPP02] [HPP05]*
  - *Functional disability in elderly people [JRR 2004]*
  - *Follow up [MCM 2012]*
  - *Urban planning [NNW05]*
  - *Dependency in severe mental illness [HARPS 2010]*
  - *Comorbidity between severe mental disease and intellectual disability [AIA2007]*
  - *Response to rehabilitation in acquired brain dammage [MedArch2008],*
  - *successfull therapies? (in press)*
  - *Quality of life perceived in patients with spinal cord injury [StudHTI 09] [ActaInfMed2009]*
  - *Profile processes in waste water treatment plant [EMS2010]*
  - *Characterization of Agitation episodes in severe mental disease [BMC Psychi 2017]*
  - *Characterization of Delta del Ebre visitors [Information and Management 2017]*
  - *Mental Health Systems in under-developed countries (in press)*
  - *Types of Borderline Personality Disorder (in press)*

Karina Gibert
Dpt. Statistics and Operation Research
Knowledge Engineering and Machine Learning Research group at Intelligent Data Science and Artificial Intelligence Research Center
Research Institute of Sience and Technology of Sustainability
Universitat Politècnica de Catalunya-BarcelonaTech (Spain)
karina.gibert@upc.edu
www.eio.upc.edu/en/homepages/karina



Are there any questions?...