Clustering Validation

K. Gibert[1]

[1]Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group at
Intelligent Data Science and Artificial Intelligence Research Center
KEMLG-@-IDEAI (UPC)
Universitat Politècnica de Catalunya, Barcelona

Karina.gibert@upc.edu
https://www.eio.upc.edu/en/homepages/karina

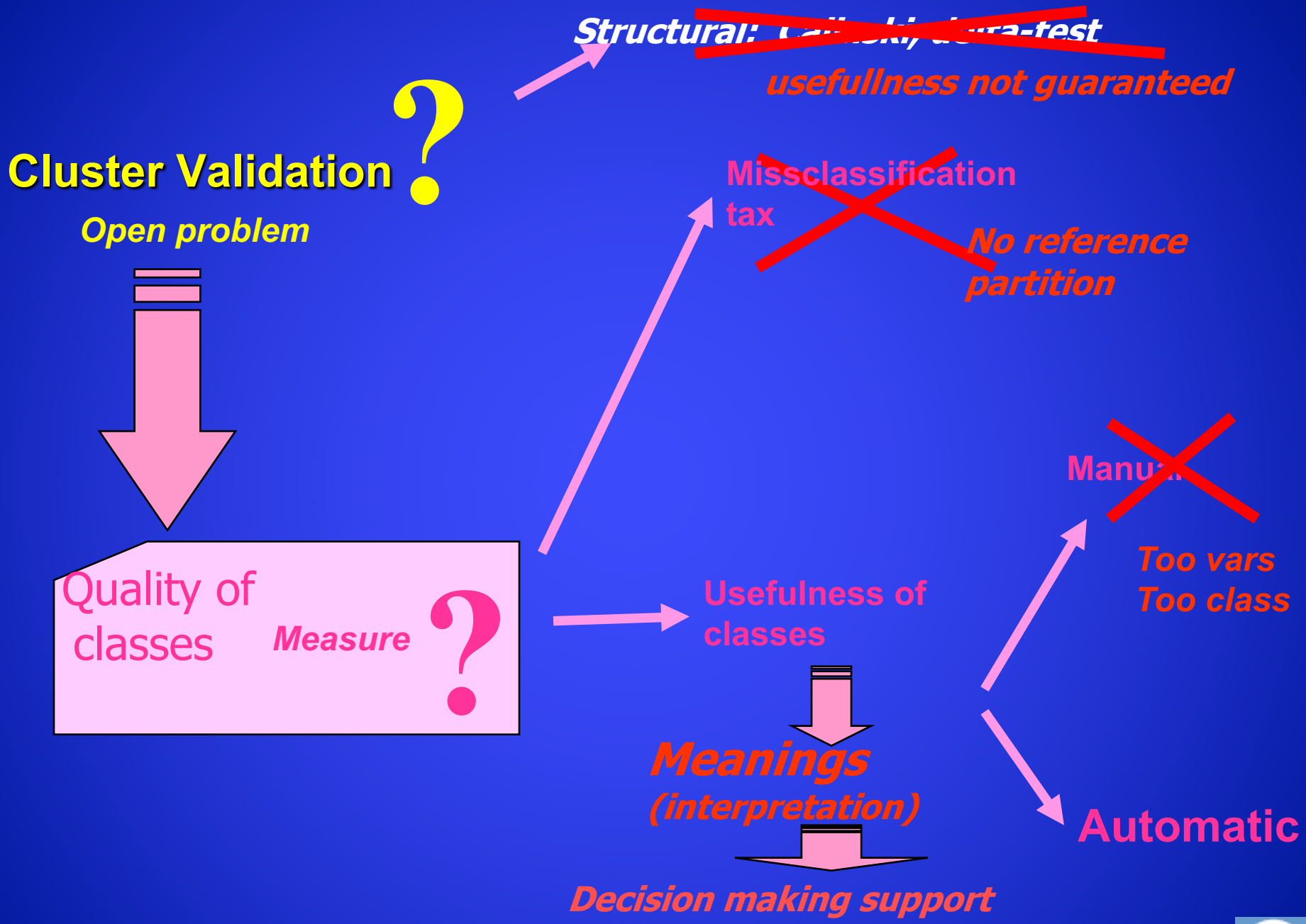
Validation

Properties of a good clustering:

- Small number of clusters
  - Large coverage → good generality
- Big cluster descriptions
  - More features → more inferential power
- Minimal or no overlap between clusters
  - More distinct clusters → better defined concepts

Post-processing in clustering

- Validation:
  - Structural (still open problem)
  - Conceptual: Intepretation of the classes


Structural:
usefullness not guaranteed
Cluster Validation
?
Open problem
Missclassification tax
No reference partition
Quality of classes
Measure
?
Usefulness of classes
Too vars
Too class
Meanings
(interpretation)
Automatic
Decision making support

Validation Criteria

- Extrinsic criteria:
  - Ask to the expert
  - True class of some data points (semisupervised)

  - F-measure (Calinski-Harabasz)
  - Consistency mearuse (or purity measure)
    - RAND INDEX, adjusted rand index
    - All assume existence of reference classes

- Intrinsic criteria:
  - Compacity/connexity/separation
  - Linear combinarion (validity index SD)
  - Non linear combinarion Davies-bouldin/Dunn-like/Silhouette

  This criteria are not of upper level than the one used to optimize. There is no reason to justify its use as a validation criteria

Validation Criteria

- Use a stability criterion
- If the cluster keeps stable over variations on the clustering parameterization it must be a true cluster
- Multiple clustering:
  - Consensus clustering
    - Make several clusterings with different parameters
    - Match the classes among them
    - Find common structures

Structural validation

- Traditional evaluation:

$$ClusteringQuality :\uparrow \frac{InterClusterDistance}{IntraClusterDistance}$$

- Cluster validity indexes
  - Calinski-Harabasz
  - Inertia Ratios
  - Entropy
  - Jonyer

- No known evaluation for hierarchical clusterings
  - Most hierarchical evaluations are anecdotal

Structural validation
Calinski Harabasz Index (1974)

Redundant in Hierarchical methods

$$CH_k = \frac{B_k / (k-1)}{W_k / (n-k)}$$

Cluster is better for high values

, being k the number of clusters,

, B_k the between classes variability: $$B_k = \sum_{C \in P} n_C \mathrm{d}(\bar{i}_C, \bar{i})^2$$

, W_k the within classes variability: $$W_k = \sum_{C \in P} \sum_{\mathrm{i,i' \in C}} \mathrm{d(i,i')}^2$$

, $\bar{i}_C$ the centroid of the cluster C, $\bar{i}$ the centroid of the whole dataset

[Calinski74] Caliński T., Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 1974;3(1):1-27.

Other coefficients [Gibert 06]

- Inertia ratios:

$$S_c^2 = \frac{\sum_{\forall i \in c} d(x_i, \bar{x}_c)^2}{n_c - 1}$$

$$S_p^2 = \frac{\sum_{\forall c} (n_c - 1) S_c^2}{n - \xi}$$

$$S_\xi^2 = \frac{\sum_{\forall c} d(\bar{x}_c, \bar{x})^2}{n - \xi}$$

$$F = \frac{S_\xi^2}{S_p^2}$$

- Entropy: How much random X is wrt Y

$$I(X_1,\ldots,X_k,Y) = \sum_{x_1} \ldots \sum_{x_k} \sum_{y_i} \Pr(x_1,\ldots,x_k,y_i) \log \frac{\Pr(x_1,\ldots,x_k,y_i)}{\Pr(x_i,\ldots,x_k)\Pr(y_i)}$$

Entropy

$$Entropy = -\sum_{i}^{k} \frac{|c_i|}{n} * \log\left(\frac{|c_i|}{n}\right)$$

n: num.objects
k: num.clusters
c_i: cluster i
$|c_i|$: number of objects of cluster i

$$I(X_1,\ldots,X_k,Y) = \sum_{x_1}\ldots\sum_{x_k}\sum_{y_i}\Pr(x_1,\ldots,x_k,y_i)\log\frac{\Pr(x_1,\ldots,x_k,y_i)}{\Pr(x_i,\ldots,x_k)\Pr(y_i)}$$

Heuristic for Hierarchical Clustering *[Jonyer, U. Texas]*

$$CQ_C = \frac{\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\sum_{k=1}^{|H_i|}\sum_{l=1}^{|H_j|}\frac{distance(H_{i,k}, H_{j,l})}{\left\|\max_{size}(H_{i,k}, H_{j,l})\right\|}}{\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}(|H_i| * |H_j|)} + \sum_{i=1}^{c} CQ_{H_i}$$

Big clusters: bigger distance between disjoint clusters

Overlap: less overlap → bigger distance

Few clusters: averaging comparisons

Structural Validation
Cluster Validity Indexes

Index	Meaning	Optimal Value
Dunn	Separation vs Compactness	Maximize
Pearson version of Hubert's Gamma coefficient (Pearson)	Correlation	1
Average of Silhouette Width	Compactness vs Separation to the nearest cluster	1
Calinski – Harabasz (CH)	Separation vs Compactness	Maximize
Average Distance Between	Separation	Maximize
Minimum Cluster Separation	Separation	Maximize
Separation Index	Separation	Maximize
Average Distance Within	Compactness	Minimize
Goodman and Kruskal's G3	Compactness	Minimize
Maximum Cluster Diameter	Compactness	Minimize

Índice de Davies-Bouldin (BDI) (1979)

El índice de Davies-Bouldin (DBI) se basa en relacionar la dispersión dentro de los clústeres (intra-clúster) y la separación entre clústeres (inter-clúster)

$$S_i = \frac{1}{k} \sum_{i=1; i \neq j}^{k} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

$k \equiv Número\ de\ clusters$

$\sigma_i \equiv distancia\ promedio\ entre\ cada\ punto\ en\ el\ cúster\ i\ y\ el\ centroide$

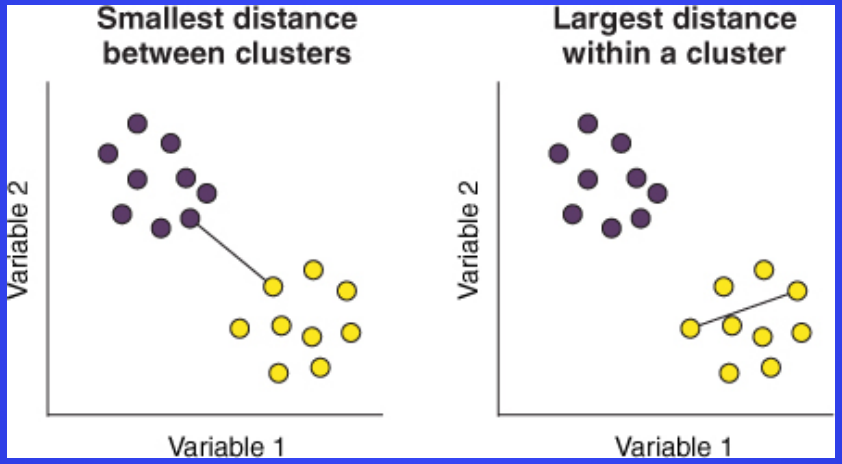
$\sigma_j \equiv distancia\ promedio\ entre\ cada\ punto\ en\ el\ cúster\ j\ y\ el\ centroide$

$d(c_i, c_j) \equiv distancia\ ente\ los\ centroides\ de\ los\ 2\ clústers$

Valores pequeños indican clústeres compactos y centroides bien separados los unos de los otros

Índice de Dunn (1974)

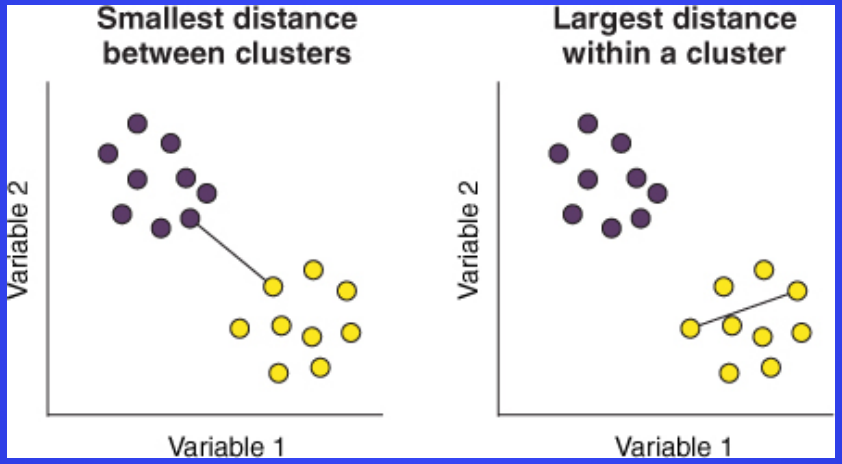
El índice de Dunn es la relación entre la distancia más pequeña entre las observaciones que no están en el mismo grupo y la distancia más grande dentro del grupo.



El índice de Dunn toma valores entre 0 y infinito y el objetivo es maximizar dicho valor

Índice de Dunn (1974)

El índice de Dunn (validación interna) es la relación entre la distancia más pequeña entre las observaciones que no están en el mismo grupo y la distancia más grande dentro del grupo.



El índice de Dunn toma valores entre 0 y infinito y el objetivo es maximizar dicho valor

Índice de Dunn (1974)

$$D = \frac{\min_{C,C' \in P} \delta(C,C')}{\max_{C \in P} \Delta_C}$$

$$\delta = \min_{C,C' \in P} \delta_{C,C'} \qquad \Delta = \max_{C \in P} \Delta_C$$

$$\delta_{C,C'} = \min_{i \in C, i' \notin C} d(i,i') \qquad \Delta_C = \max_{i,i' \in C} \mathrm{d(i,i')}$$

[Dunn74] J. C. Dunn: Well Separated Clusters and Optimal Fuzzy Partitions, Journal of Cybernetica, Vol. 4, pp. 95-104, 1974

[Halkidi01] Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3), 107-145.

[Brun07] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, E.R. Dougherty, Model- based evaluation of clustering validation measures, Pattern Recognition 40 (2007) 807–824.

Índice de Dunn-Like (1998)

El índice de Dunn – Like (Bezdek, 1998) es una de las generalizaciones del índice Dunn (Dunn, 1974) propuesto por Bezdek y Pal (1998). Intenta identificar grupos comptactos y bien separados pero buscando robustez en la búsqueda de distancias (Halkidi 2001)

$$D = \frac{\min_{C,C' \in P} \overline{\delta(C,C')}}{\max_{C \in P} \overline{\Delta_C}}$$

$$\overline{\delta(C,C')} = \frac{\sum_{i \in C, i' \in C'} d(i,i')}{n_C n_{C'}}$$

$$\overline{\Delta_C} = \frac{\sum_{i,i' \in C} d(i,i')}{n_C (n_C - 1)}$$

[Bezdek98]J. C. Bezdek and N. R. Pal. Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics PART B: CYBERNET- ICS , 28, no. 3:301-315, 1998.

[Dunn74] J. C. Dunn: Well Separated Clusters and Optimal Fuzzy Partitions, Journal of Cybernetica, Vol. 4, pp. 95-104, 1974

[Halkidi01] Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3), 107-145.

Índice de Silhouette (Rousseu 1987)

El coeficiente de Silueta (Rousseu 1987) es una métrica para evaluar la calidad del agrupamiento obtenido con algoritmos de segmentación identificando el número mínimo de grupos óptimos que se necesitan para segmentar un conjunto de datos.

OBJETIVO: Valorar el número óptimo de agrupamientos

$$s(i) = \frac{b - a}{\max\{a, b\}}$$

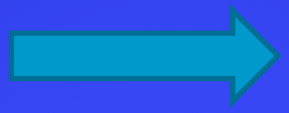
$$S(i) = \begin{cases} 1 - \frac{a}{b} & si\ a < b \\ 0 & si\ a = b \\ \frac{b}{a} - 1 & si\ a > b \end{cases}$$

$a \equiv promedio\ de\ las\ distancias\ de\ la\ observación\ i\ con\ las\ demás\ observaciones\ del\ clúster$

$b \equiv distancia\ mínima\ a\ otro\ clúster\ (segunda\ opción\ o\ vecindad)$

Índice de Silhouette (Rousseu 1987)

$$s(i) = \frac{b - a}{\max\{a, b\}}$$



$$S(i) = \begin{cases} 1 - \frac{a}{b} & si\ a < b \\ 0 & si\ a = b \\ \frac{b}{a} - 1 & si\ a > b \end{cases}$$

El coeficiente de Silueta es un valor que varía entre -1 y 1

Su interpretación es:

$S(i) \approx 1 \equiv La\ observación\ i\ está\ bien\ asignada\ a\ su\ clúster$

$S(i) \approx 0 \equiv La\ observación\ i\ está\ entre\ dos\ clusters$

$S(i) \approx -1 \equiv La\ observación\ i\ está\ mal\ asignada\ a\ su\ clúster$

Índice de Silhouette (Rousseu 1987)

$$s(i) = \frac{b - a}{\max\{a, b\}}$$

$$S(i) = \begin{cases} 1 - \frac{a}{b} & si\ a < b \\ 0 & si\ a = b \\ \frac{b}{a} - 1 & si\ a > b \end{cases}$$

Se puede calcular el coeficiente de Silueta como:

$$S = \frac{1}{n} \sum s(i)$$

Coeficiente generalmente más alto para grupos convexos, bien separados y con una densidad alta

Cluster interpretation

- Graphical: Class panel graph

  Trafic lights panel

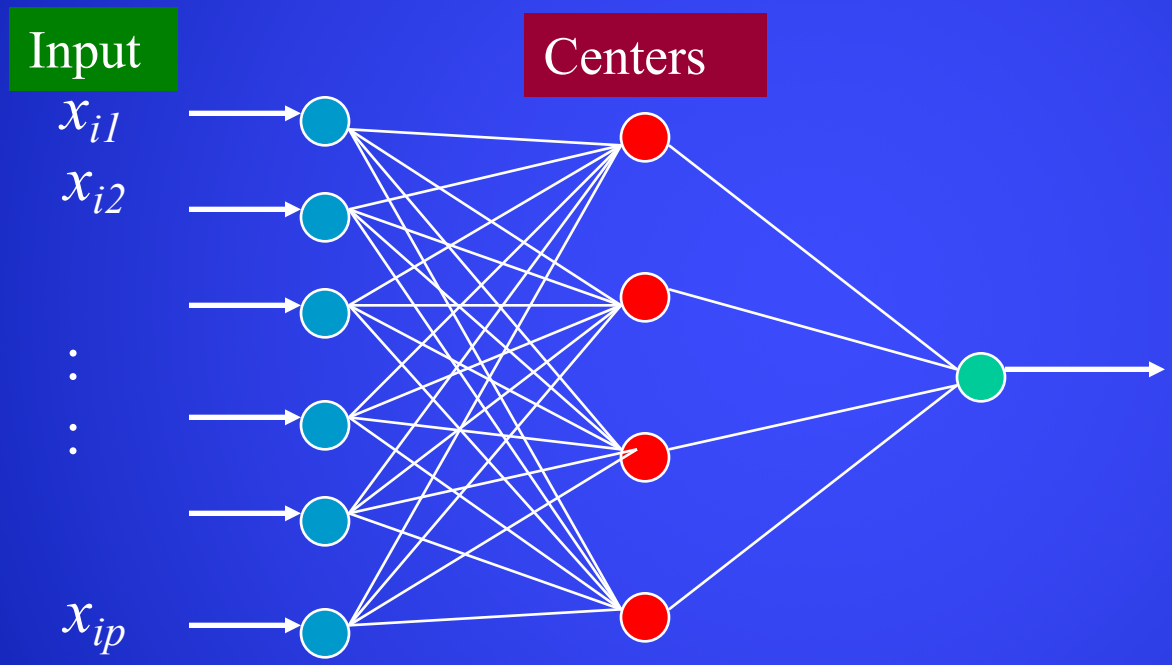
- Numerical: Test each variable against the classes: ANOVA, Kruskal-Wallis, profiling

- Conceptual: find distinctive class characters

  (CCEC)

Assignment of new individuals

- Define rules to assign new individuals
- Compute the distance of the new individual to each class centroid



crisp $x_i \rightarrow C_h$ argmin $d(x_i, C_h)$

soft $x_i \rightarrow C_h$ random draw $f(d(x_i, C_h), h=1\ldots K)$