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:

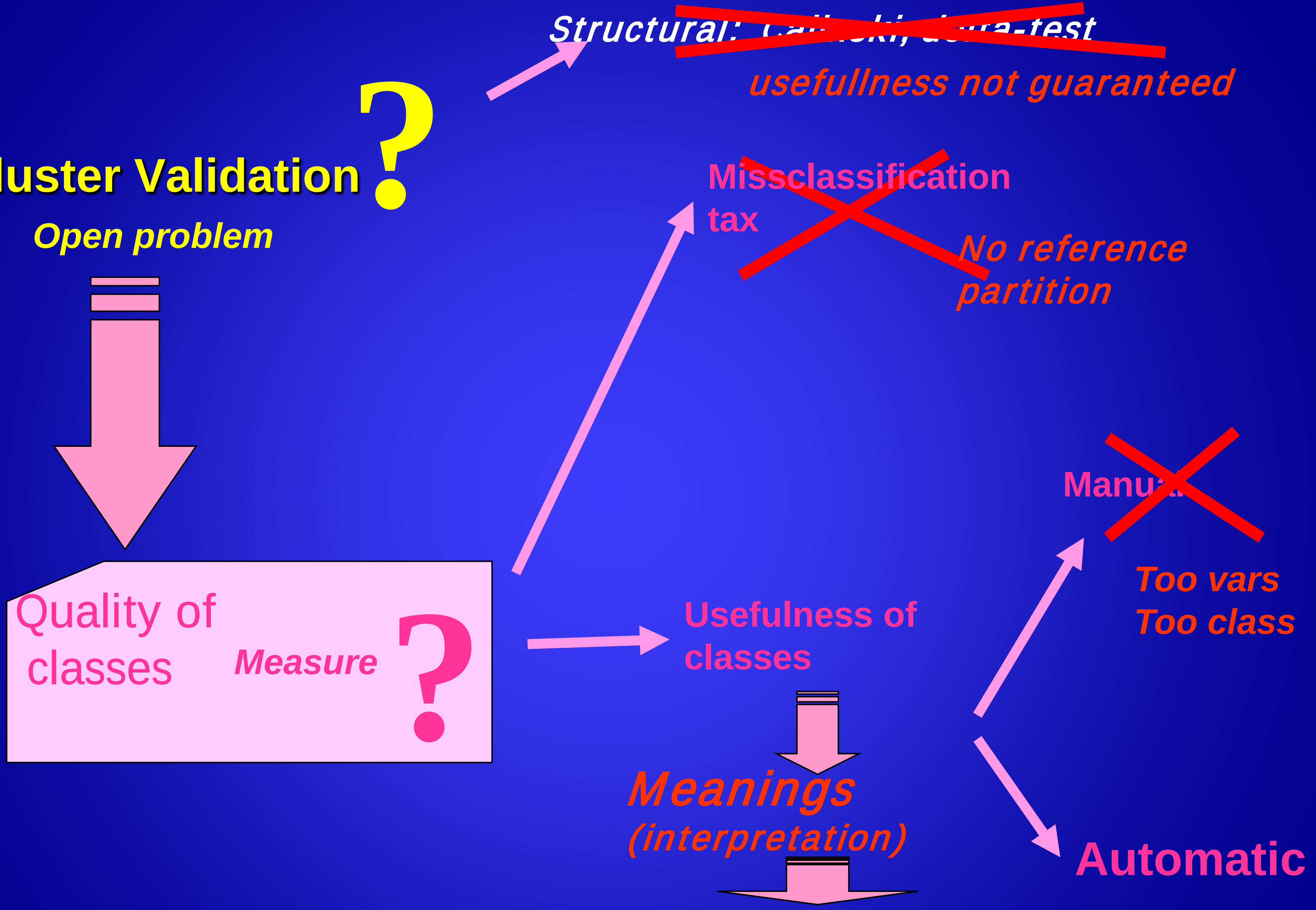
- Small number of clusters
  - Large coverage → good generality
- Big cluster descriptions
  - More features → more inferential power
- Minimal or no overlap between clusters
  - More distinct clusters → better defined concepts

Post-processing in clustering

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

Cluster Validation ?

Open problem



Quality of classes *Measure* ?

Structural: ~~Calinski, delta-test~~

usefullness not guaranteed

~~Missclassification tax~~

No reference partition

Usefulness of classes

Meanings

(interpretation)

Decision making support

~~Manual~~

Too vars
Too class

Automatic

©K. Gibert

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:

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

- 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: $B_k = \sum_{C \in P} n_C \, \mathrm{d}(\bar{\imath}_C, \bar{\imath})^2$

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

, $\bar{\imath}_C$ the centroid of the cluster C, $\bar{\imath}$ 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]*

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

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

Davies-Bouldin

- L’utilitza molt per a triar el millor numero de classes
- el menor minimitza dist intra clusters I maximitza la entre:

•e Davies-Bouldin Index evaluates intra-cluster similarity and inter-cluster differences. If you consider these to be good criteria, go for the Davies-Bouldin.

Dunn Index (D)

Dunn Index (*D*) is a cluster validity index for crisp clustering proposed in Dunn (1974) [Dunn74]. It attempts to identify “compact and well separated clusters”[Halkidi02]. If a data set contains well-separated clusters, the distances among the clusters are usually larger than the diameters of the clusters. We present a formulation from [Brun07]

$$D = \frac{\min\limits_{C,C'\in P} \delta(C,C')}{\max\limits_{C\in P} \Delta_C}$$

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

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

Optimal Value: Maximum

[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.

Dunn Index (D)

Dunn Index (*D*) is a cluster validity index for crisp clustering proposed in Dunn (1974) [Dunn74]. It attempts to identify “compact and well separated clusters”[Halkidi02]. If a data set contains well-separated clusters, the distances among the clusters are usually larger than the diameters of the clusters. We present a formulation from [Brun07]

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

$$\delta = \min_{C,C'\in P}\left[\min_{i\in C, i'\notin C} d(i,i')\right]$$

$$\Delta = \max_{C\in P}\left[\max_{i,i'\in C} \mathrm{d}(\mathrm{i},\mathrm{i}')\right]$$

Optimal Value: Maximum

[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.

Dunn-like

Dunn-like index is one of the generalizations of Dunn index[Dunn74] proposed by Bezdek and Pal[Bezdek98]. It attempts to identify compact and well separated clusters. This is one generalization from all proposed in [Bedzdek98] .This version is more robust than the original Dunn Index [Halkidi01].

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

$$\overline{\delta(C,C')} = \frac{\sum_{i\in C.i'\in C'} \mathrm{d}(\mathrm{i},\mathrm{i}')}{\mathrm{n}_C \mathrm{n}_{C'}}$$

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

Optimal Value: Maximum

[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.

Silhouette Index

The Silhouettes Index [Rousseu87] provides a succinct graphical representation of how well each object lies within its cluster.

$$S = \frac{\sum_{i \in C} s(i)}{n_C} \qquad s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

a(i): average dissimilarity of *i* with all other data within the same cluster. *a(i)* shows us how well clustered *i* is to the cluster assigned (smaller value means better matching).

b(i): the lowest average dissimilarity of *i* with the data of another single cluster. The cluster with this lowest average dissimilarity is said to be the "neighbouring cluster" of *i* as it is, aside from the cluster *i* is assigned, the cluster in which *i* fits best.

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

$$b(i) = \min_{C' \neq C} \frac{\sum_{i \in C, i' \in C'} d(i,i')}{n_{C-1}}$$

Optimal Value: 1

[Rousseu87] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65

The Silhouette Index measure the distance between each data point, the centroid of the cluster it was assigned to and the closest centroid belonging to another cluster. If you consider that this is a good criterion, go for the silhouette index.
How can we say that a clustering quality measure is good? - ResearchGate. Available from: https://www.researchgate.net/post/How_can_we_say_that_a_clustering_quality_measure_is_good [accessed Dec 14, 2016].

Davies Bouldin Index (DB)

The Davies Bouldin Index [Davies79] is a cluster separation measure. The overall index is defined as the average of indices computed from each individual cluster. An individual cluster index is taken as the maximum pairwise comparison involving the cluster and the other clusters in the solution.

$$DB = \frac{1}{\xi}\sum_{C\in P}^{\xi} \max_{C',C''\neq C}\left(\frac{S_{p_C}+S_{p_{C'}}}{d_p(C,C')}\right)$$

$$d_p(C,C') = \sqrt[p]{\sum_{k=1}^{K} |\overline{x_{C_k}} - \overline{x_{C'_k}}|^p}$$

$$S_{p_C} = \sqrt[p]{\frac{\sum_{i\in C} d_p(i,i_C)^p}{n_C}}$$

Where, i_C : barycenter of the cluster C

$$i_C = (\overline{x_{C_1}},\ldots,\overline{x_{C_K}}) \qquad \overline{x_{C_k}} = \frac{\sum_{i\in C} x_i}{n_C}$$

p: Minkowski factor (p=1 Manhattan distance, for p=2 Euclidean distance)

Sp_C: dispersion measure of a cluster C (for p=1 the average distance of objects in cluster C to the barycenter of cluster C; for p=2 the standard deviation of the distance of objects in cluster C to the barycenter of cluster C)

Optimal Value: Minimum

D. L. Davies and D. W. Bouldin: Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1, No. 2, pp. 95-104, 1979

Baker and Hubert Index (variant of Goodman and Kruskal)

$$BH(k) = \frac{S^{+} - S^{-}}{S^{+} + S^{-}}$$

S^{+} :number of concordant quadruples,
S^{-}: number of discordant quadruples.
For this index, all possible quadruples (q, r, s, t) of input parameters are considered.
Let $d(i,i')$ = distance between the samples i and i'.

A quadruple is called concordant if one of the following two conditions is true:

- $d(q,r) < d(s,t)$, q and r are in the same cluster, and s and t are in different clusters.
- $d(q,r) > d(s,t)$, q and are in different clusters, and $s$and t are in the same cluster.

By contrast, a quadruple is called discordant if one of following two conditions is true:

- $d(q,r) < d(s,t)$, q and r are in different clusters, and s and t are in the same cluster.
- $d(q,r) > d(s,t)$, q and r are in the same cluster, and s and t are in different clusters.

Baker FB, Hubert LJ (1975). "Measuring the Power of Hierarchical Cluster Analysis." Journal of the American Statistical Association, 70(349), 31-38.

Modified Hubert Γ Statistic

The definition of the modified Hubert Gamma statistic is given by the equation :

$$\Gamma = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P(i,j)\cdot Q(i,j)$$

where,

$$M = \frac{N(N-1)}{2}$$

N : *number of objects*

P : Distance matrix, $P(i, j) = distance\ (x_i, x_j)$

Q : $N \times N$ matrix ,whose $Q(i, j) =$ *distance* between the centers of the clusters where x_i *and* x_j belong respectively

Hubert LJ, Arabie P (1985). "Comparing partitions". Journal of Classification, 2, 193-218

Goodman and Kruskal's Gamma coefficient (G)

$$G = \frac{N_s - N_d}{N_s + N_d},$$

N_s: the number of pairs of cases ranked in the same order on both variables

N_d: the number of pairs of cases ranked differently on the variables (discordant pairs),

Gordon A (1999). "Classification". 2nd edition. Chapman & Hall/CRC, London. ISBN 1-58488-013-9.

Hubert & Levin C-Index

$$C = \frac{S - S_{\min}}{S_{\max} - S_{\min}}$$

- S: Sum of distances over all pairs of samples from the same cluster.
- Let l be the number of those pairs. Then S_{min} is the sum of the l smallest distances if all pairs of samples are considered (i.e. if the sample can belong to different clusters).
- Similarly, S_{max} is the sum of the l largest distances out of all pairs.

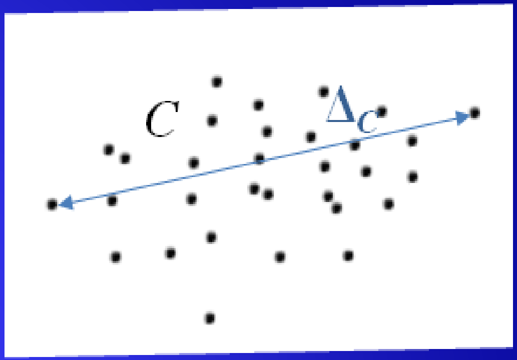
Milligan and Cooper (1985), Hubert and Levin (1976), Gordon (1999)

Hubert LJ, Levin JR (1976). "A General Statistical Framework for Assessing Categorical Clustering in Free Recall". Psychological Bulletin, 83(6), 1072-1080.

$\Delta = \max_{C \in P} \Delta_C$

Maximum Diameter

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



$$\Delta = \max_{C \in P} \Delta_C$$

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

WidestGap

$$widestgap = \max_{i=1..k}\left[\mathrm{max(heights(hierarchicalClustering\_SingleLink(D_i)))}\right]$$

- k: num.clusters
- D_i: distance matrix of cluster i

Average within distance

$$Avg.within = mean(avg.distance_i)$$

$$= \frac{\sum_{i}^{k} \sum_{x,y \in c_i} d(x,y)^2}{\sum_{i}^{k} |c_i|(|c_i|-1) / 2}$$

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

Sindex

$$sepprob = ct(ex.0.1)$$

$$sindex = \frac{\sum_{i=1}^{minsep} sort.sep(i)}{minsep}$$

$$sort.sep = sort(sep)$$

$$sep(i) = \min_{x \in c_i \& y \notin c_i} (d(x, y))$$

$$minsep = floor(n * sepprob)$$

Hennig, C. (2013) How many bee species? A case study in determining the number of clusters. To appear in Proceedings of GfKl-2012, Hildesheim.

Modified Hubert Γ statistic

Dunn Index

Silhoutte index

Hubert & Levine C-Index G3

Goodman and Kruskal's Gamma coefficient (G)

Baker and Hubert index (G2)

Cluster interpretation

- Graphical: Class panel graph

  Trafic lights panel

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

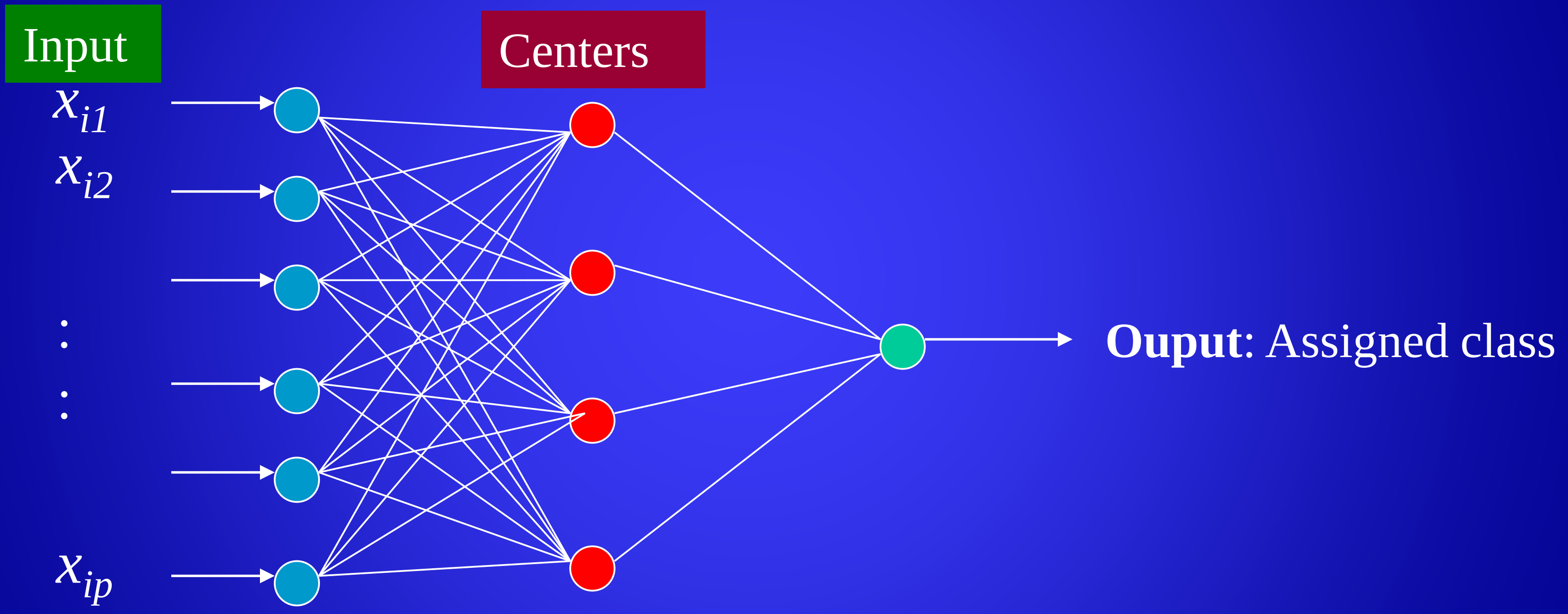
n not too large

- Conceptual: find distinctive class characters (CCEC)

Care with classical classifiers

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 $f(d(x_i, C_h), h=1\ldots\mathrm{K})$