CLUSTERING

CLUSTER VALIDITY – PART II

Fabio Stella

Associate Professor

c/o Department of Informatics, Systems and Communication

University of Milano-Bicocca

Cluster Validity

The following concepts will be introduced:

✓ **Internal or Unsupervised indices**

- Cohesion
- Separation
- Silhouette coefficient
- Cophenetic Correlation coefficient

Many internal measures or indices of cluster validity for partitional clustering schemes are based on the notions of **Cohesion** or **Separation**.

In general, we can consider expressing the overall cluster validity for a set of *K* clusters

$$C_1, \ldots, C_K$$

as a weighted sum of the validity of individual clusters C_i:

$$\text{overall validity} = \sum_{i=1}^{K} w_i \cdot \text{validity}(C_i)$$

where the validity function can be cohesion, separation or any combination of them.

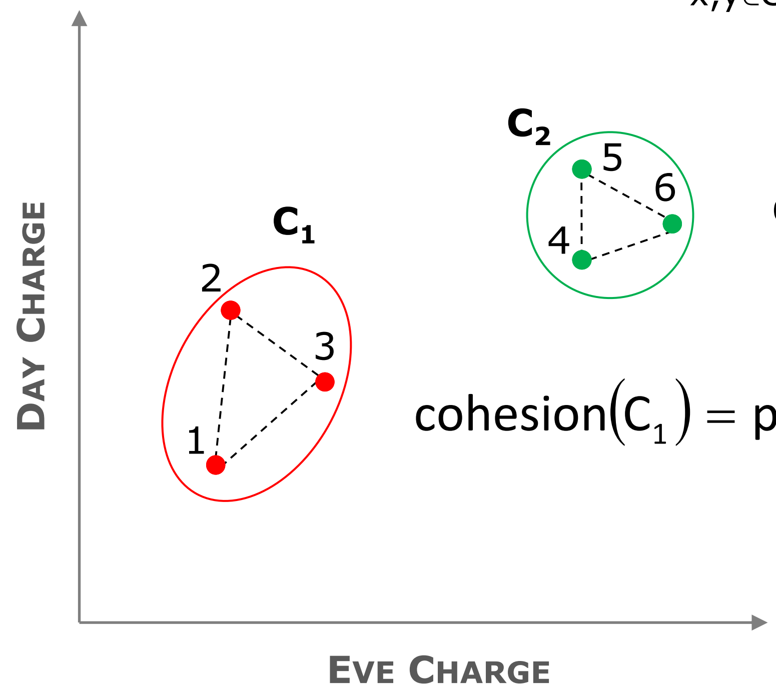
The weights will vary depending on the clustering validity measure. In some cases they are set to 1 or are the cardinality of the corresponding cluster, while in other cases they reflect a more complicated property, such as the square root of the cohesion.

- ✓ *Validity* = **cohesion** higher values are better
- ✓ *Validity* = **separation** lower values are better

Cluster Validity: Internal or Unsupervised 2

For **graph-based clusters**, the **cohesion of a cluster** can be defined as the sum of the weights of the links in the proximity graph that connect points within the cluster.

$$\text{cohesion}(C_i) = \sum_{x,y \in C_i} \text{proximity}(x,y) = \sum_{x,y \in C_i} \text{similarity}(x,y)$$



$$\text{cohesion}(C_2) = \text{proximity}(4,5) + \text{proximity}(4,6) + \text{proximity}(5,6)$$

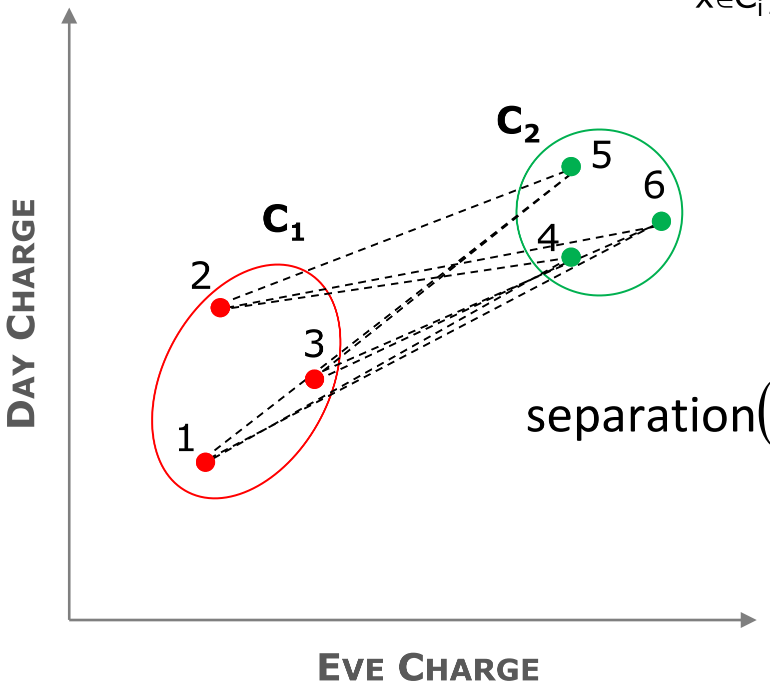
$$\text{cohesion}(C_1) = \text{proximity}(1,2) + \text{proximity}(1,3) + \text{proximity}(2,3)$$

Therefore, **cohesion** and **similarity** are maximized when **dissimilarity/distance** are minimized.

When considering the attributes' space, it is useful to recall that **similarity** is inversely proportional to **dissimilarity/distance**.

For **GRAPH-BASED CLUSTERS**, **SEPARATION BETWEEN TWO CLUSTERS** can be measured by the sum of weights of the links from points in one cluster to points in the other cluster.

$$separation(C_i,C_j)=\sum_{x\in C_i,y\in C_j} proximity(x,y)=\sum_{x\in C_i,y\in C_j} similarity(x,y)$$



Therefore, **SEPARATION** and **SIMILARITY** are minimized when **DISSIMILARITY/DISTANCE** are maximized.

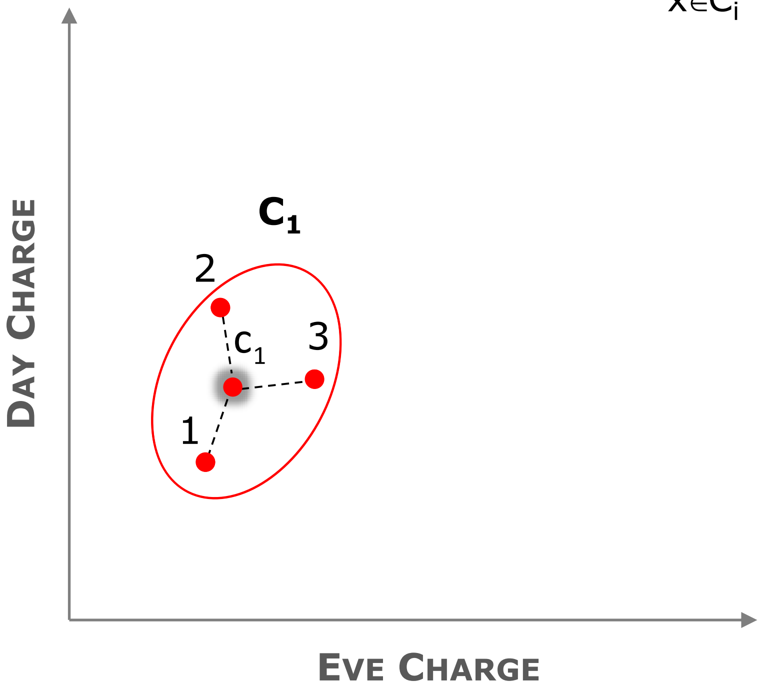
$$\begin{aligned} separation(C_1,C_2) = {} & proximity(1,4)+proximity(1,5)+proximity(1,6)+ \\ & +proximity(2,4)+proximity(2,5)+proximity(2,6)+ \\ & +proximity(3,4)+proximity(3,5)+proximity(3,6) \end{aligned}$$

When considering the attributes' space, it is useful to recall that **SIMILARITY** is inversely proportional to **DISSIMILARITY/DISTANCE**.

For **prototype-based clusters**, the **cohesion of a cluster** can be defined as the sum of the proximities with respect to the prototype (centroid or medoid) of the cluster.

$$\text{cohesion}(C_i) = \sum_{x \in C_i} \text{proximity}(x, c_i) = \sum_{x \in C_i} \text{similarity}(x, c_i)$$

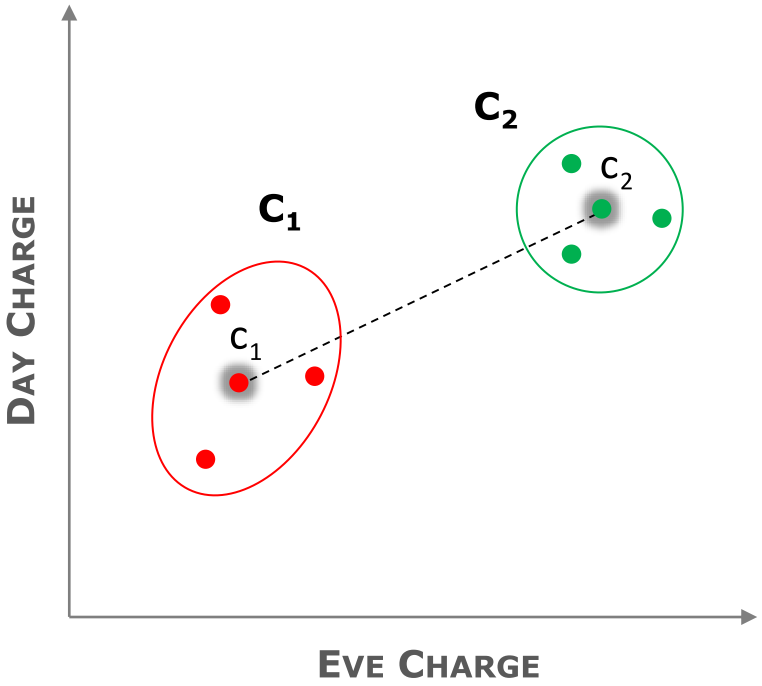
Centroid or Medoid of cluster C_i



$$\text{cohesion}(C_1) = \sum_{x \in C_1} \text{proximity}(x, c_1) = \text{proximity}(1, c_1) + \text{proximity}(2, c_1) + \text{proximity}(3, c_1)$$

For **PROTOTYPE-BASED CLUSTERS**, the **SEPARATION BETWEEN TWO CLUSTERS** can be measured by the proximity of the two clusters prototypes.

$$\mathrm{separation}(C_i, C_j) = \mathrm{proximity}(c_i, c_j) = \mathrm{similarity}(c_i, c_j)$$

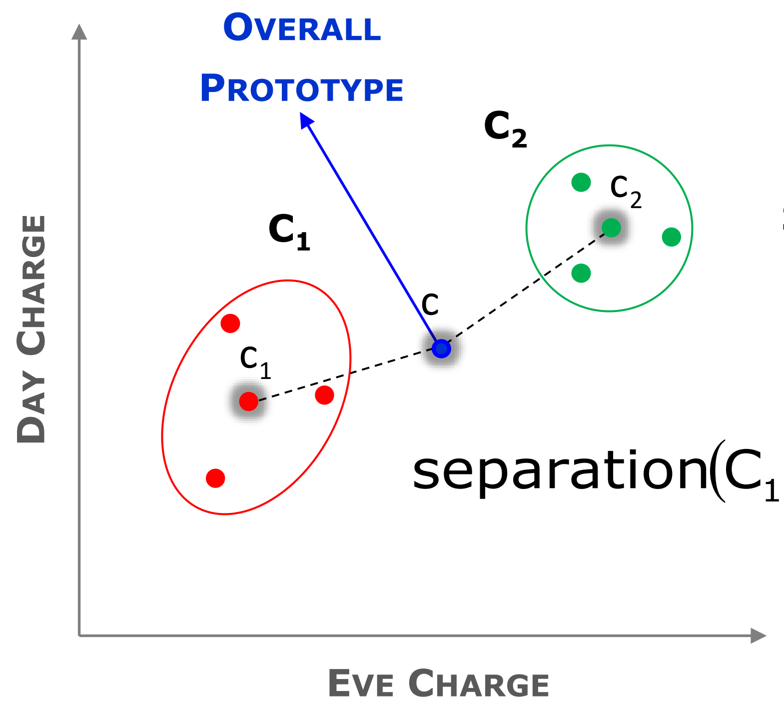


$$\mathrm{separation}(C_1, C_2) = \mathrm{proximity}(c_1, c_2)$$

For **PROTOTYPE-BASED CLUSTERS**, the **SEPARATION BETWEEN TWO CLUSTERS** can be measured by the proximity of the two clusters prototypes.

$$\text{separation}(C_i) = \text{proximity}(c_i, c) = \text{similarity}(c_i, c)$$

OVERALL PROTOTYPE



$$\text{separation}(C_2) = \text{proximity}(c_2, c)$$

$$\text{separation}(C_1) = \text{proximity}(c_1, c)$$

The previous definitions of cluster cohesion and separation offer simple and well-defined measures of cluster validity that can be combined into an overall measure of cluster validity by using a weighted sum

$$\text{overall validity} = \sum_{i=1}^{K} w_i \cdot \text{validity}(C_i)$$

However, we need to decide what weights w_i to use. Indeed, the weights used can vary widely, although typically they express a measure of the cluster size. Some examples are:

Cluster Measure	Cluster Weight	Type
$\text{cohesion}(C_i) = \sum_{x,y \in C_i} \text{proximity}(x, y)$	$\frac{1}{m_i}$	Graph-Based cohesion
$\text{cohesion}(C_i) = \sum_{x \in C_i} \text{proximity}(x, c_i)$	1	Prototype-Based cohesion
$\text{separation}(C_i) = \text{proximity}(c_i, c)$	m_i	Prototype-Based separation

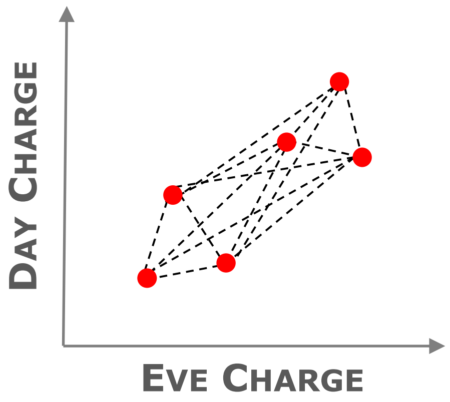
Potentially any unsupervised measure of cluster validity can be used as an objective function for a clustering algorithm and vice versa.

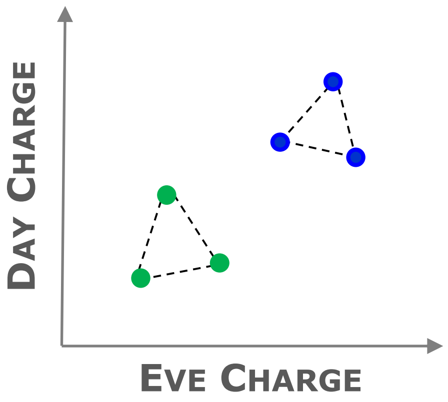
So far, we focused on cohesion and separation for the overall evaluation of a group of clusters. Many of these measures of cluster validity also can be used to evaluate individual clusters and objects (records).

We could rank individual clusters according to their specific value of cluster validity, i.e., cluster cohesion or separation. A cluster that has a high value of cohesion can be considered better than a cluster that has a lower value.

This information is useful to **IMPROVE THE QUALITY OF THE CLUSTER ANALYSIS PROCESS**.

Cluster not very cohesive → **split into several clusters**





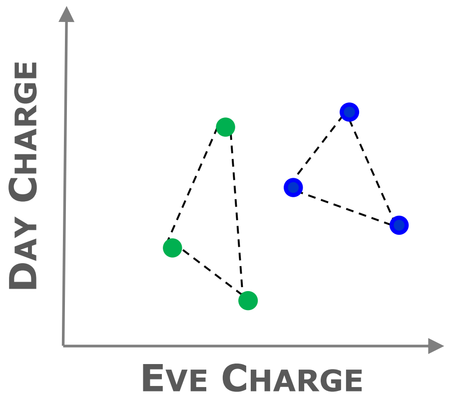
CLUSTER VALIDITY: INTERNAL OR UNSUPERVISED

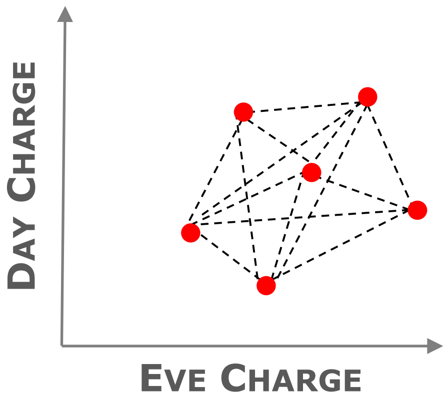
So far, we focused on cohesion and separation for the overall evaluation of a group of clusters. Many of these measures of cluster validity also can be used to evaluate individual clusters and objects (records).

We could rank individual clusters according to their specific value of cluster validity, i.e., cluster cohesion or separation. A cluster that has a high value of cohesion can be considered better than a cluster that has a lower value.

This information is useful to **IMPROVE THE QUALITY OF THE CLUSTER ANALYSIS PROCESS**.

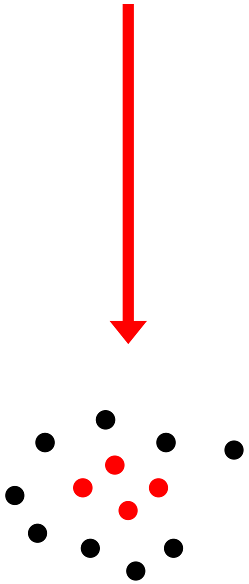
Two clusters are relatively cohesive but not well separated → **merge them into a single cluster**





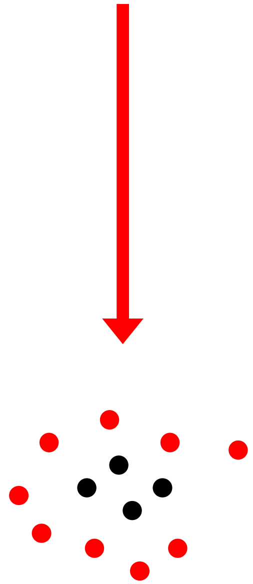
We can also evaluate the objects within a cluster in terms of their contribution to the overall cohesion or separation of the cluster.

Objects that contribute more to the overall cohesion and separation of a cluster



near the interior of the cluster

Objects that contribute less to the overall cohesion and separation of a cluster



near the edge of the cluster

The **SILHOUETTE COEFFICIENT** is a cluster evaluation measure which exploits the concepts of interior and edge of a cluster to evaluate data points, clusters and the entire set of clusters.

The **SILHOUETTE COEFFICIENT** combines cohesion and separation and for the i^{th} object (record) it is defined as follows

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, +1]$$

where

a_i average distance of the i^{th} object to all other objects in its cluster

b_i minimum of the average distances of the i^{th} object to all the objects in each given cluster different from the cluster to which the i^{th} object belongs to

NEGATIVE SILHOUETTE COEFFICIENT means that the average distance to points in its cluster (a_i) is greater than the minimum average distance to points in another cluster (b_i).

We want that the **SILHOUETTE COEFFICIENT** is positive ($a_i < b_i$), and for a_i to be as close to 0 as possible, since the **SILHOUETTE COEFFICIENT** assumes its maximum value of 1 when $a_i = 0$.

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, +1]$$

where

a_i average distance of the i^{th} object to all other objects in its cluster

b_i minimum of the average distances of the i^{th} object to all the objects in each given cluster different from the cluster to which the i^{th} object belongs to

Negative Silhouette Coefficient means that the average distance to points in its cluster (a_i) is greater than the minimum average distance to points in another cluster (b_i).

We want that the **Silhouette Coefficient** is positive ($a_i < b_i$), and for a_i to be as close to 0 as possible, since the **Silhouette Coefficient** assumes its maximum value of 1 when $a_i = 0$.

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, +1]$$

We can compute the **Average Silhouette Coefficient of a cluster** by simply taking the average of the Silhouette Coefficients of data points (records) belonging to the considered cluster.

An overall measure of goodness of a clustering can be obtained by computing the **Average Silhouette Coefficient of all points**.

The **SILHOUETTE COEFFICIENT** is defined for partitional clustering while for hierarchical clustering a different evaluation measure is used; **COPHENETIC CORRELATION COEFFICIENT**.

It measures the degree of similarity between the **PROXIMITY MATRIX P** and the **COPHENETIC MATRIX Q** whose elements record the proximity level where pairs of data points are grouped in the same cluster for the first time.

The value of the **COPHENETIC CORRELATION COEFFICIENT** lies in the range of [-1,1], and the index value close to 1 indicates a significant similarity between **P** and **Q** and a good fit of hierarchy to the data.

However, for average linkage, even large values of the **COPHENETIC CORRELATION COEFFICIENT** cannot assure sufficient similarity between the two matrices.