cluster Validation (cluster Evaluation)

Evaluation of Clustering

⇒ As of now, we have applied various clustering algorithms on data set. Next thing is, how to evaluate whether the clustering results are good or not.

⇒ Major tasks of clustering evaluation includes

① o Assessing cluster tendency :-

In this, we assess whether data set consists of non-random structure in the data. Clustering analysis is meaningful when there is a non-random structure in the data. This can be done using Hopkin's statistic as in (1).

... in a dataset.

Hopkin's statistic :-

1. we generate p points that are randomly uniformly distributed across the data space; {i.e. each point has the same probability of being included in the sample}

2. Also, sample p actual data points from dataset D.

3. Let u_i be nearest neighbour distance of the artificially generated points and w_i be nearest neighbour distance of the sample of points from the original dataset.

$$\text{Hopkin statistic}(H) = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

- if both points have same nearest neighbour distance, H is nearly 0.5 [redacted]

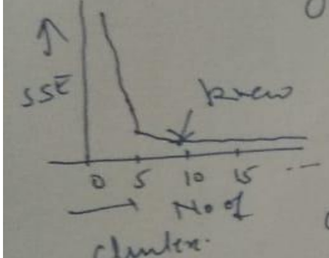
$H = 0 \Rightarrow$ High clustering data

$H = 1 \Rightarrow$ randomly distributed

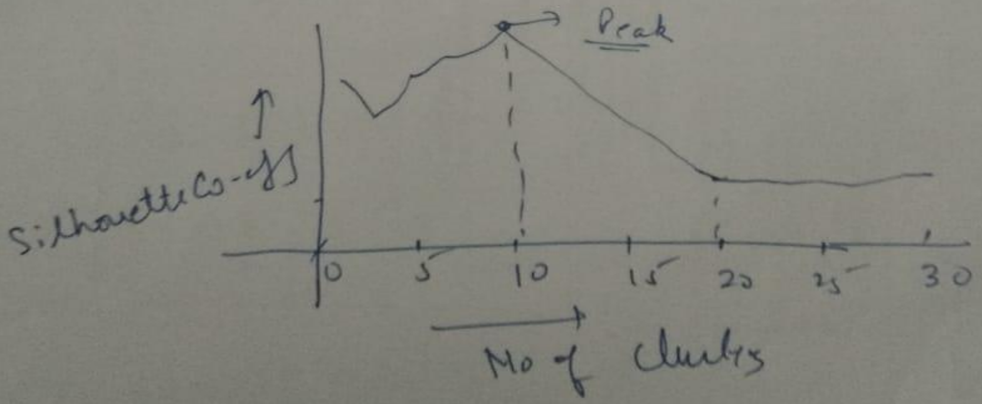
(2) Determining the number of clusters in a dataset.

It is desirable to estimate this number even before a clustering algorithm is used to derive detailed clusters.

Various unsupervised evaluation measures can be used to approximately determine the correct or natural no of clusters.



- SSE (Sum of Squared Errors)
  i.e. Plot a graph B/w SSE and no of clusters and find knee in SSE.
- Silhouette co-efficient
  i.e. Plot a graph B/w this co-eff. and no of clusters and find a distinct peak in the silhouette co-eff.



(3) ○ Measuring Cluster Quality :-

→ How good clustering method resulted into clusters ?

→ Two methods :-

- ○ Extrinsic Methods (when ground truth is known)
- ○ Intrinsic Methods (When no ground truth is known)

Intrinsic Methods :-

- ○ Ground truth is not known
- ○ Hence, this is also known as unsupervised cluster evaluation.

In general, overall cluster validity for a set of K clusters is a weighted sum of the validity of individual clusters.

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

Validity function can be cohesion, separation or some combination of them. Weights may vary on validity measure.

In some cases, as weights are simple 1 or

→ Size of the cluster or

→ Square root of cohesion and others.

Higher value of cohesion and lower value of separation are better.

(i) Cohesion (Validity function)

→ Let weight = 1

→ $\text{Cohesion}(C_i) = \sum_{x \in C_i} \text{Proximity}(x, c_i)$

c_i is the centroid of cluster C_i

x belongs to cluster C_i

$\therefore \text{Cohesion}(C_i) = \text{dist}(x_1, c_i) + \text{dist}(x_2, c_i) + \text{dist}(x_3, c_i) + \dots \text{dist}(x_n, c_i)$

$(x_1, x_2 \dots x_n) \in C_i$

$$\therefore \boxed{\text{Overall validity} = \sum_{i=1}^{K} w_i \, \text{Cohesion}(C_i)}$$

$= \text{Cohesion}(C_1) + \text{Cohesion}(C_2) + \dots \text{Cohesion}(C_K)$

K = No of clusters.

+ Note:- if proximity measure is Euclidean distance, cohesion is the cluster SSE (Sum of squared errors).

(ii) Separation (Validity function)

Separation $(\mathcal{C}_i, \mathcal{C}_j)$ = Proximity (c_i, c_j) —— (1)

c_i & c_j are centroid of $\underline{\mathcal{C}_i}$ and $\underline{\mathcal{C}_j}$ ($\mathcal{C}_i$ and $\mathcal{C}_j$ are clusters)

Separation $(\mathcal{C}_i)$ = Proximity (c_i, c) —— (2)

c_i = centroid of cluster $\mathcal{C}_i$

c = overall centroid (mean of all points)

(iii) Silhouette Co-eff: (Combination of (i) & (ii))

The following steps are taken to compute this co-eff for an individual point.

(a) For the i^{th} object, calculate the average distance to all other objects in its cluster. Call this value a_i.

(b) For the i^{th} object and any cluster not containing the object, calculate object's average distance to all the objects in the given cluster. Find the minimum such value with respect to all clusters, call this value b_i.

(c) for the i^{th} object, silhouette co-eff is defined as

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad \text{———} \quad (3)$$

$$\boxed{-1 \leq S_i \leq 1}$$

→ negative value is undesirable

→ Positive value is desirable.

→ overall measure of goodness of clustering can be obtained by computing the <u>average silhouette co-eff of all points</u>.

$$\text{overall silhouette co-eff } (S) = \frac{1}{n}\sum_{i=1}^{n} S_i$$

n = no of data points

S_i = Calculated from equation 3.

(II) Hierarchical Clustering :-

→ These algorithms can be compared by calculating $\boxed{\text{Cophenetic}}$ Correlation co-efficient

→ Cophenetic Coefficient gives the correlation B/w the distance matrix (dissimilarity matrix) and cophenetic distance matrix.

→ Cophenetic matrix is computed using cophenetic distance which is the proximity at which an agglomerative hierarchical clustering technique puts the objects in same cluster for the first time.

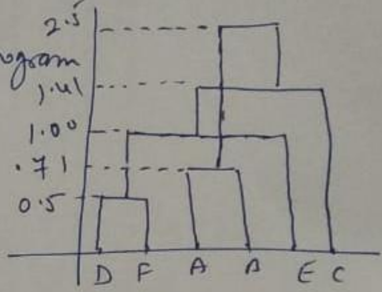
→ Eg Let the smallest distance B/w two objects (clusters) that are merged is 0.1, then all points in one cluster have a cophenetic distance of 0.1 with respect to points in other cluster.

→ In cophenetic distance matrix, entries are the cophenetic distances B/w each pair of objects.

Eg Let the distance matrix (dissimilarity matrix)

	A	B	C	D	E	F
A	0					
B	.71	0				
C	5.66	4.95	0			
D	3.61	2.92	2.24	0		
E	4.24	3.54	1.41	1.00	0	
F	3.2	2.5	2.5	0.5	1.12	0

and dendrogram



Now compute the cophenetic matrix as follows

(i) ⓐ Firstly as (D, F) are merged with value 0.5, hence write the entry at (D, F) as 0.5

ⓑ Similarly then, (A,B) are merged so place 0.71 for (B, A)

	A	B	C	D	E	F
A	0					
B	.71	0				
C	2.5	2.5	0			
D	2.5	2.5	1.41	0		
E	2.5	2.5	1.41	1	0	
F	2.5	2.5	1.41	.5	1	0

ⓒ (D, F) and E at 1.0 hence DE & FE = 1.0

ⓓ ((D, F), E) and C at 1.41, hence DC = FC = EC = 1.41

ⓔ (((D, F), E), C) and (A, B) at 2.5, hence

DA = DB = 2.5 & CA = CB = 2.5

FA = FB = 2.5

EA = EB = 2.5

(ii) Find the correlation co-eff B/w this matrix (Cophenetic Matrix) and dissimilarity matrix.

(X) dissimilarity	CP (Y)
① 0.71	0.71
② 5.66	2.50
③ 3.61	2.50
④ 4.24	2.50
⑤ 3.20	2.50
⑥ 4.95	2.50
⑦ 2.92	2.50
⑧ 3.54	2.50
⑨ 2.50	2.50
⑩ 2.24	~~2.50~~ 1.41
⑪ 1.41	1.41
⑫ 2.50	1.00
⑬ 1.00	0.50
⑮ 0.50	1.00

Pearson Co-eff (r)

$$= \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}}$$

$\bar{x}$ = mean of X variable

$\bar{y}$ = mean of Y variable.

(iii) Higher the value, better the clustering is.

Extrinsic Methods for cluster evaluation :-

⇒ Ground truth is known (i.e. we know the class labels for data objects)

⇒ When we have class labels, why to do clustering?

Ans :- Clustering analysis helps to automatically do the classification process that was done manually.

⇒ Hence, we do the cluster evaluation using supervised measures.

⇒ These measures calculate the degree of correspondence B/w cluster labels and class labels. (ground truth).

Approaches

- classification Oriented (entropy, Precision, Recall, F-measure)
- Similarity Oriented (Related to similarity Measures like binary data, Jaccard co-eff, Rand statistics

A) Similarity-Oriented Measure:

- In this method, ideal cluster similarity matrix and ideal class similarity matrix (as shown in Table 8.10 and Table 8.11) are found and compared.

- Then, a simple matching coefficient such as **Rand Statistic , Jaccard coefficient** is computed.

$$\text{Rand statistic} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

$$\text{Jaccard coefficient} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

f_{00} = number of pairs of objects having a different class and a different cluster
f_{01} = number of pairs of objects having a different class and the same cluster
f_{10} = number of pairs of objects having the same class and a different cluster
f_{11} = number of pairs of objects having the same class and the same cluster

	Same Cluster	Different Cluster
Same Class	f_{11}	f_{10}
Different Class	f_{01}	f_{00}

- Example of Matrices:

Table 8.10. Ideal cluster similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	1	0	0
p2	1	1	1	0	0
p3	1	1	1	0	0
p4	0	0	0	1	1
p5	0	0	0	1	1

Table 8.11. Ideal class similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	0	0	0
p2	1	1	0	0	0
p3	0	0	1	1	1
p4	0	0	1	1	1
p5	0	0	1	1	1

(i) Ideal cluster similarity matrix has 1 in ij^{th} entry if two objects, i and j, are in the same cluster, else "0".

(ii) Ideal class similarity matrix is defined with respect to class labels, which has 1 in ij^{th} entry if two objects, i & j, belong to the same class, and "0" otherwise.

as shown in Table 8.10 and 8.11 for the following example :-

Let $D = \{P_1, P_2, P_3, P_4, P_5\}$ has two clusters as $C_1 = \{P_1, P_2\}$

$C_2 = \{P_3, P_4, P_5\}$

<u>Calculation</u> of <u>Rand statistics</u> (Matching Co-eff)

→ It is easy to define two way contigency table to determine whether pair of objects are in the same class and same clusters.

1 (same)
0 (different)

	Same cluster	Different cluster
Same class	f_{11}	f_{10}
Different class	f_{01}	f_{00}

Table 8.10. Ideal cluster similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	1	0	0
p2	1	1	1	0	0
p3	1	1	1	0	0
p4	0	0	0	1	1
p5	0	0	0	1	1

Table 8.11. Ideal class similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	0	0	0
p2	1	1	0	0	0
p3	0	0	1	1	1
p4	0	0	1	1	1
p5	0	0	1	1	1

Note: f_{ij} ($i \Rightarrow$ class, $j \Rightarrow$ cluster)

We need to compute these quantities for all pair of distinct objects. (i.e. total no of pair of objects $= m\cdot(m-1)/2$, where m is no of objects in dataset).

$f_{11} \Rightarrow (P_1 P_2), (P_4 P_5)$ [2]

$f_{00} \Rightarrow (P_1 P_4), (P_1 P_5), (P_2 P_4), (P_2 P_5)$ [4]

$f_{01} \Rightarrow (P_1 P_3), (P_2 P_3)$ [2]

$f_{10} \Rightarrow (P_3 P_4), (P_3 P_5)$ [2]

$$\therefore \text{Rand Statistic} = \frac{(f_{00}+f_{11})}{(f_{00}+f_{11})+(f_{01}+f_{10})}$$

$$= \frac{6}{6+(2+2)} = \frac{6}{10} = 0.6$$

$$\therefore \text{Jaccard coefficient} = \frac{f_{11}}{(f_{01}+f_{10})+f_{11}} = \frac{2}{(2+2)+2}$$

$$= \frac{2}{6} = \frac{1}{3} = 0.33$$

Classification - Oriented Approaches

These measures evaluate the extent to which two objects that are in the same class. In this context, predicted class labels are the cluster labels.

(i) Entropy :-

→ The degree to which each cluster consists of objects of a single class.

→ entropy of each cluster i is :-

$$e_i = -\sum_{j=1}^{L} P_{ij} \log_2 P_{ij}$$, where L is no of classes.

where $P_{ij} = \left(\frac{m_{ij}}{m_i}\right)$

⇒ m_i = no of objects in cluster i.

⇒ P_{ij} = the probability that a member of cluster i belong to class j

⇒ m_{ij} = no of objects of class j in cluster i.

→ Total entropy of a set of clusters is calculated as:

$$e = \sum_{i=1}^{K} \frac{m_i}{m}(e_i)$$ where K is no of clusters & m is the total no of points.

e_i = entropy of i^{th} cluster.

(ii) Purity :-

→ The extent to which a cluster contains objects of a single class.

→ Purity of cluster $i = P_i = \max_j (P_{ij})$

$$\text{overall purity} = \sum_{i=1}^{K} \left(\frac{m_i}{m}\right)(P_i)$$

(iii) Precision :-

→ The fraction of a cluster that consists of objects of a specified class.

→ The precision of cluster i with respect to class j = Precision $(i, j) = P_{ij} = \left(\frac{m_{ij}}{m_i}\right)$

(iv) Recall :-

→ The extent to which a cluster contains all objects of a specific class.

→ The recall of cluster i with respect to class j = recall $(i, j) = \left(\frac{m_{ij}}{m_j}\right)$

m_j = no of objects in class j

(v) F-measure :-

→ Combination of precision and recall.

→ $$F(i,j) = \frac{2 \times Precision(i,j) * Recall(i,j)}{Precision(i,j) + Recall(i,j)}$$

⇓

F-measure of cluster i with respect to class j.

Let us take an example shown in following table which is a result of K-means clustering of 3204 news articles consisting of 6 classes from India. No of clusters are 6.

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports
1	3	5	40	506	96	27
2	4	7	280	29	39	2
3	1	1	1	7	4	671
4	10	162	3	119	73	2
5	331	22	5	70	13	23
6	5	358	12	212	48	13
Total	354	555	341	943	273	738

- Ideally, each cluster contains documents from only one class.
- But in reality, each cluster contains documents from many classes.
- Many clusters contain documents primarily from just one class.