Clustering

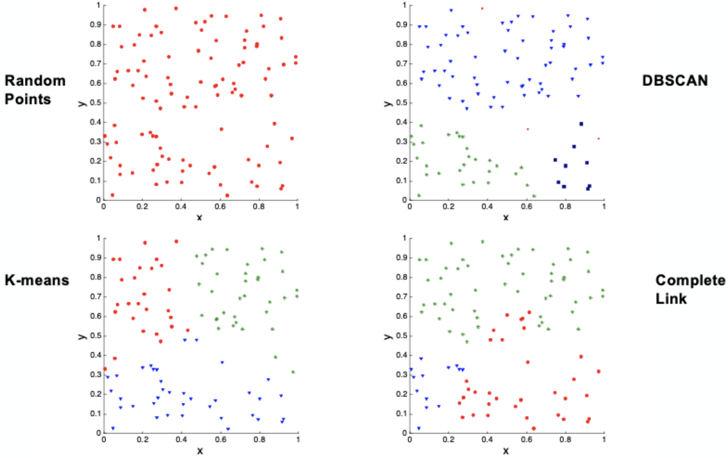
Cluster Evaluation

Huiping Cao

Cluster Validity

- For supervised classification we have a variety of measures to evaluate how good our model is. E.g., accuracy, precision, recall
- For cluster analysis, the analogous question is **how to evaluate the "goodness" of the resulting clusters**?
- Cluster analysis is conducted as a part of an exploratory data analysis. "Clusters are in the eye of the beholder"!
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters

Clusters found in Random Data



Different Aspects of Cluster Validation

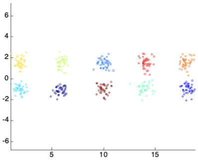
- Determining **the clustering tendency** of a set of data, i.e., distinguishing whether non-random structure actually exists in the data. (purely unsupervised)
- **Comparing** the results of a cluster analysis **to externally known** results, e.g., to externally given class labels. (supervised/unsupervised)
- Evaluating how well the results of a cluster analysis fit the data **without reference to external** information. (Use only the data) (purely unsupervised)
- **Comparing** the results of **two different sets of cluster analyses** to determine which is better.
- Determining the **"correct" number** of clusters. (purely unsupervised)

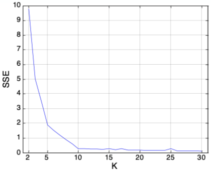
Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - **Internal Index**: Used to measure the goodness of a clustering structure without respect to external information.
    - Sum of Squared Error (SSE)
  - **External Index**: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - **Relative Index**: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as **criteria** instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

Internal Measures: SSE

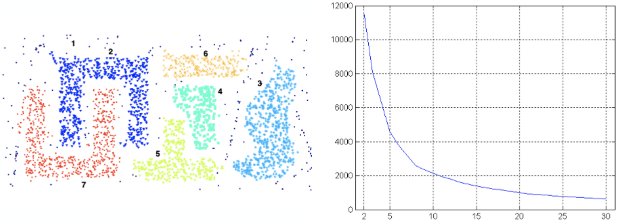
- Clusters in more complicated figures are not well separated
- Internal Index: Used to measure the goodness of a clustering structure without respect to external information
  - SSE
- **SSE is good for comparing two**clusterings or two clusters (average SSE).
- Can also be used to **estimate the number of clusters**





Internal Measures: SSE (cont.)

- SSE curve for a more complicated data set



SSE of clusters found using K-means

Internal Measures: Cohesion and Separation

- **Cluster Cohesion**: Measures how closely related are objects in a cluster
  - Example: SSE
- **Cluster Separation**: Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
  - Cohesion is measured by the **within cluster sum of squares (WSS)**

$$SSE = WSS = \sum_{i} \sum_{x \in C_i} dist(x, c_i)^2$$

    c_i is the centroid of cluster C_i

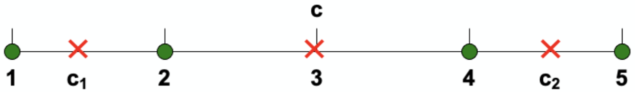
  - Separation is measured by the **between cluster sum of squares**

$$BSS = \sum_{i} |C_i| dist(c_i, c)^2$$

    c is the overall mean. $|C_i|$ is the number of points in cluster C_i.

Internal Measures: Cohesion and Separation (cont.)

- Example:

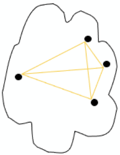


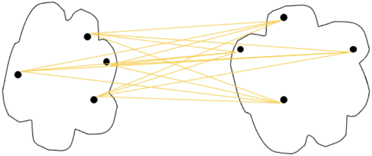
- K=1 cluster
  $SSE = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$
  $BSS = 4 \times (3-3)^2 = 0$
  $Total = 10 + 0 = 10$

- K=2 cluster
  $SSE = (1-1.5)^2 + (2-.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$
  $BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$
  $Total = 1 + 9 = 10$

Internal Measures: Cohesion and Separation

- A **proximity graph-based approach** can also be used for cohesion and separation.
  - **Cluster cohesion** is the sum of the weight of all links **within** a cluster.
  - **Cluster separation** is the sum of the weights between nodes in the cluster and nodes outside the cluster.

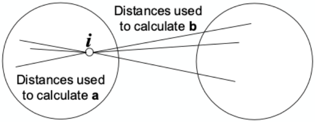

cohesion


separation

Internal Measures: Silhouette Coefficient

- **Silhouette coefficient** combines ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, **i**
  - Calculate a = average distance of **i** to the points in its cluster
  - Calculate b= min (average distance of **i** to points in another cluster)
  - The silhouette coefficient for a point is then given by

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

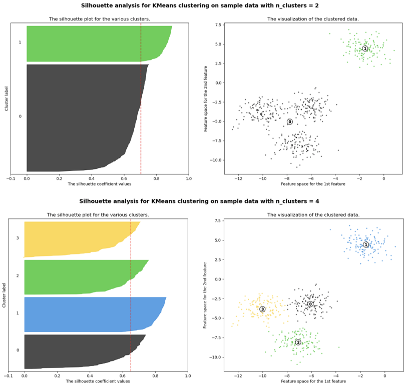


Internal Measures: Silhouette Coefficient (Cont.)

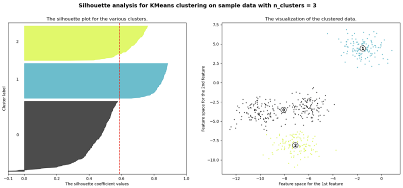
- Silhouette coefficient **ranges between -1 and 1**.
  - A negative value is undesirable because this corresponds to a case in which *a* is greater than *b*. Negative values indicate that those samples might have been assigned to the wrong cluster.
  - A positive value is desired. "+1" indicate that the sample is far away from the neighboring clusters.
  - A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters.
- We can compute the **average Silhouette coefficients of a cluster** by simply taking the average of the silhouette coefficients of points belong to the cluster.
- An **overall measure of the goodness of a clustering** can be obtained by computing the average silhouette coefficient of all points.

Use Silhouette Coefficient to Determine the Number of Clusters

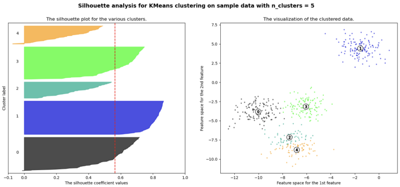
- A bad pick if there are clusters with below average silhouette scores and if there are wide fluctuations in the size of the silhouette plots.
  - It is bad to pick K=3, 5, 6.
  - Silhouette analysis is more ambivalent in deciding between 2 and 4.



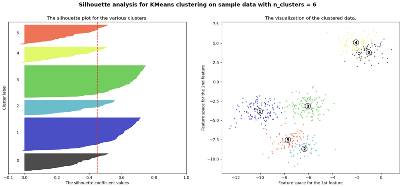
Use Silhouette Coefficient to Determine the Number of Clusters (cont.)



$K=3$



$K=5$



$K=6$

Use Silhouette Coefficient to Determine the Number of Clusters (cont.)

- The thickness of the silhouette plot can show the cluster size.
  - The silhouette plot of Cluster 0 for $K = 2$, is bigger in size owing to the grouping of the 3 sub clusters into one big cluster.
  - When $K = 4$, all the plots are more or less of similar thickness and hence are of similar sizes as can be also verified from the labelled scatter plot on the right.

Internal Measures: Silhouette Coefficient (Cont.)

- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
- Selecting the number of clusters with silhouette analysis on KMeans clustering.
  https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

External Measures of Cluster Validity: Entropy and Purity

- **Entropy**:
  - For each cluster, the **class distribution** of the data is calculated first, i.e., for cluster j, we compute $p_{i,j}$, the probability that a member of cluster j belongs to class i.

    $$p_{ij} = \frac{m_{ij}}{m_j}$$

    where m_j is the number of values in cluster C_j, and $m_{i,j}$ is the number of values of class i in cluster j.
  - The **entropy of each cluster** j is calculated using the standard formula

    $$entropy_j = \sum_{i=1}^{L} p_{ij} log(p_{ij})$$

    where L is the number of classes.
  - The **total entropy for a set of clusters** is calculated as the sum of the entropies of each cluster weighted by the size of each cluster. I.e.,

    $$entropy = \sum_{i=1}^{K} \frac{m_i}{m} entropy_i$$

    Where m_i is the size of cluster C_i, K is the total number of clusters, and m is the total number of data points.

External Measures of Cluster Validity: Entropy and Purity (cont.)

- **Purity**:
  - The purity of cluster C_j is given by

$$purity_j = max(p_{i,j})$$

  - The overall purity of a clustering is

$$purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_i$$

where K is the number of clusters.

External Measures of Cluster Validity: Entropy and Purity (cont.)

K-means Clustering Results for LA Document Data Set

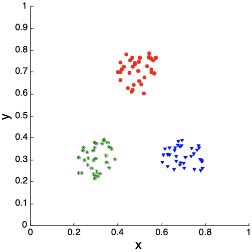
Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Measuring Cluster Validity Via Correlation

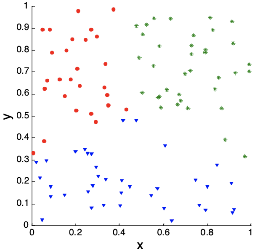
- Two **matrices**
  - **Proximity** Matrix
  - Ideal **Similarity** Matrix
    - One row and one column for each data point
    - An entry is 1 if the associated pair of points belong to the same cluster
    - An entry is 0 if the associated pair of points belongs to different clusters
- Compute the **correlation between the two matrices**
  - Since the matrices are symmetric, only the correlation between $\frac{n \cdot (n-1)}{2}$ entries needs to be calculated.
- High **correlation** indicates that points that belong to the same cluster are close to each other.
- Not a good measure for some **density or contiguity** - based clusters.

Measuring Cluster Validity Via Correlation (cont.)

- Correlation of **ideal similarity** and **proximity** matrices for the K-means clusterings of the following two data sets.



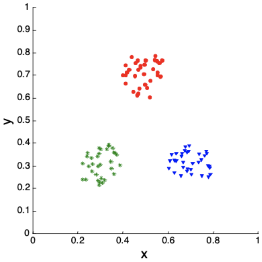
Corr = 0.9235

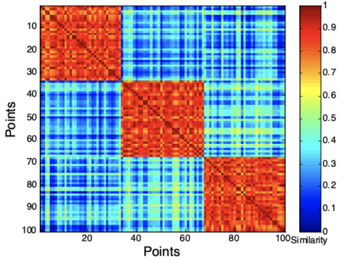


Corr = 0.5810

Using Similarity Matrix for Cluster Validation

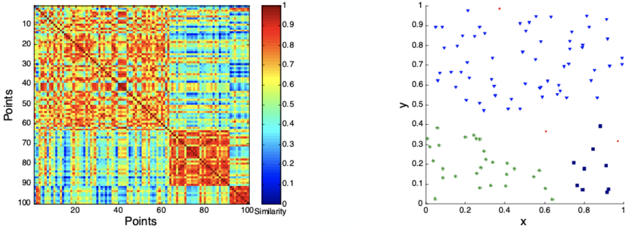
- Order the similarity matrix with respect to cluster labels and inspect visually.





Using Similarity Matrix for Cluster Validation (cont.)

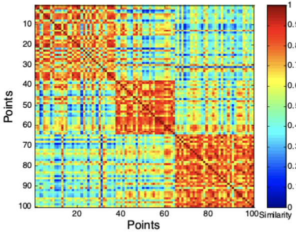
- Clusters in random data are not so crisp.

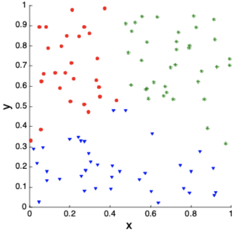


DBSCAN

Using Similarity Matrix for Cluster Validation (cont.)

- Clusters in random data are not so crisp.

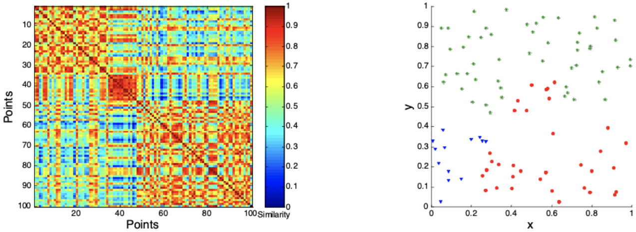




K-means

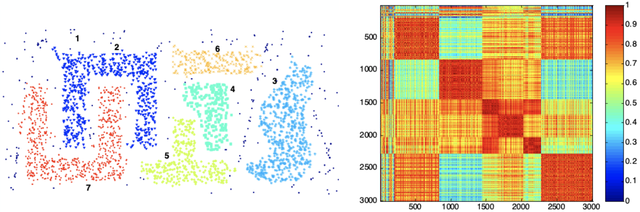
Using Similarity Matrix for Cluster Validation (cont.)

- Clusters in random data are not so crisp.



Complete Link

Using Similarity Matrix for Cluster Validation - DBSCAN results



DBSCAN

Final Comment on Cluster Validity

- "*The validation of clustering structures is the* **most difficult and frustrating** *part of cluster analysis.*
  Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."
  – Algorithms for Clustering Data, Jain and Dubes

References

- Chapter 7: Introduction to Data Mining (2nd Edition) by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar