Chapter 10. Cluster Analysis: Evaluation

Meng Jiang

Data Science

Clustering Evaluation

- Cluster Quality
  - Subjective…
  - Quantitative and objective.
  - Evaluating the goodness of a given cluster
- Cluster Stability
  - Evaluating the sensitivity of the clustering result to tunable parameters, e.g., #clusters, MinPts
- Cluster Tendency
  - Are there clusters in the data?
  - Evaluating the existence of clusters in the data

Evaluating Cluster Tendency

- Assess the suitability of clustering
  - If non-random structure exists in the data by measuring the probability that the data is generated by a uniform data distribution
- "Can I reject a hypothesis that the data are all generated from a random process that does not have cluster structure?"
  - Tests of hypotheses of randomness
- Test spatial randomness by statistic test: Hopkins Static
  - Given a dataset D regarded as a sample of a random variable x, determine how far away x is from being uniformly distributed in the data space

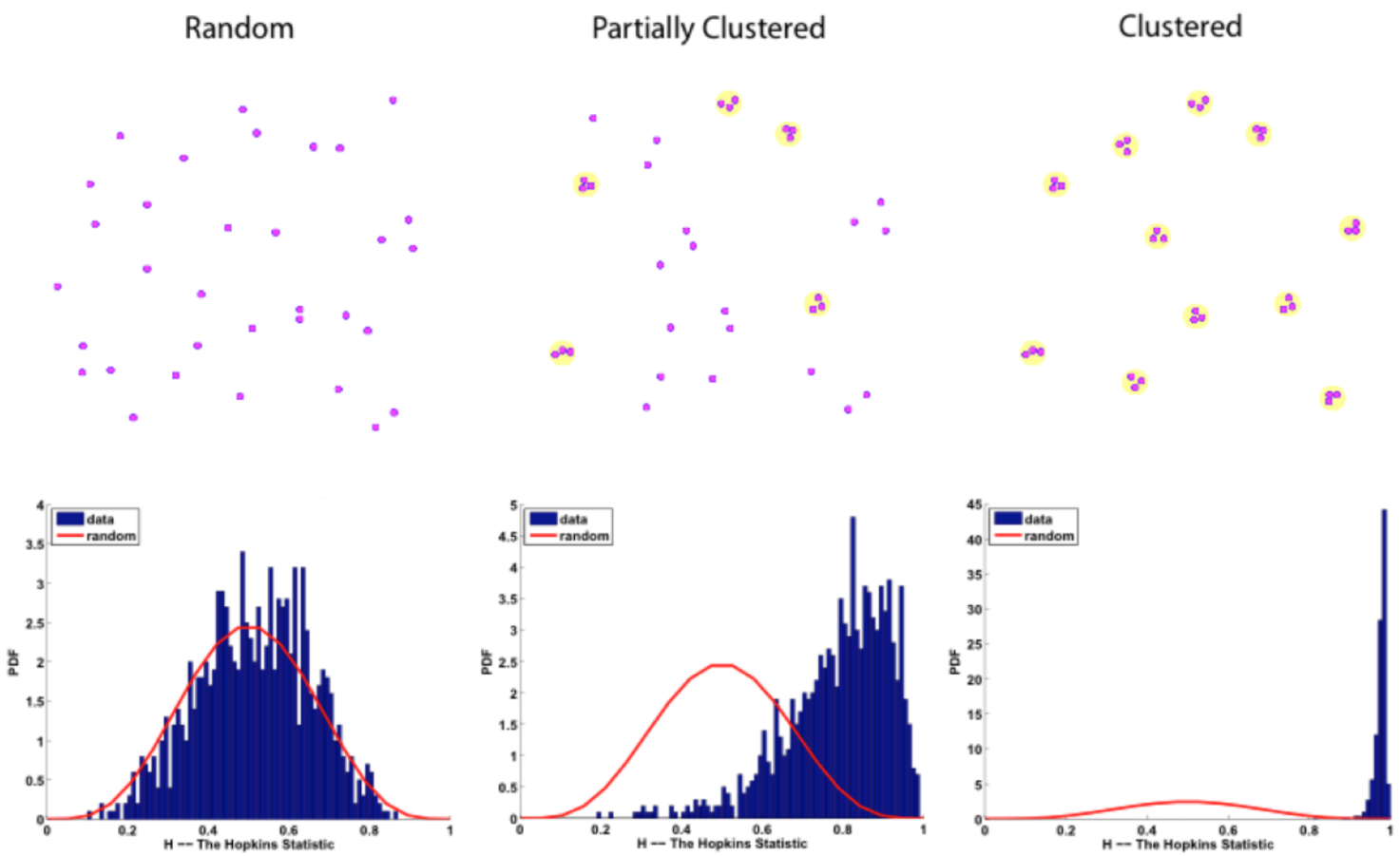
Evaluating Cluster Tendency (cont.)

- Sample *n* points, $p_1, \ldots, p_n$, uniformly from D. That is, each point in D has the same probability of being included in the sample. For each p_i, find its nearest neighbor in D – $\{p_i\}$: $x_i = min\{dist(p_i, v)\}$ where *v* in D and $v \neq p_i$
- Sample *n* points, $q_1, \ldots, q_n$, uniformly randomly distributed data points in the data space. For each q_i, find its nearest neighbor in D – $\{q_i\}$: $y_i = min\{dist(q_i, v)\}$ where *v* in D and $v \neq q_i$
- Calculate the Hopkins Statistic:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

- If D is uniformly distributed, $\sum x_i$ and $\sum y_i$ will be close to each other and H is close to 0.5 (randomness).
- If D is well-defined clustered data, H is larger than 0.5, close to 1.

Hopkins Statistic



Cluster Stability

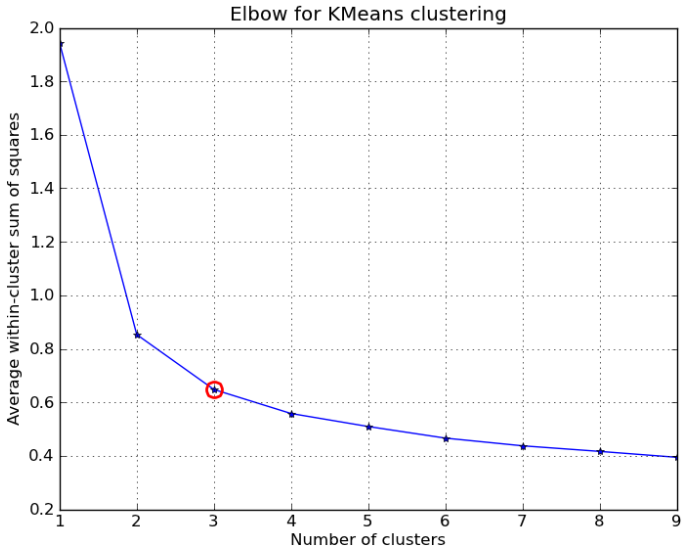
- How many clusters?
- Ad hoc: for *n* points, guess $\sqrt{\frac{n}{2}}$ for the number of clusters
- A bit unsatisfying…
- Recall clustering squared error for k clusters:

$$\min \; E_k^2 = \sum_{i=1}^{k} e_i^2 \qquad e_i^2 = \sum_{j=1}^{n_i} \left\| \vec{x}_j^{(i)} - \vec{m}^{(i)} \right\|^2$$

Cluster Stability (cont.)

- "elbow" method: plot a cluster validity index like clustering error versus number of clusters k and choose the number k* that shows a "corner" in the curve
  - Tension between "more clusters -> smaller SSE"

    and "more clusters -> flat SSE"
  - Don't look for the minimum of the curve, because it is at k* = N



Cluster Quality

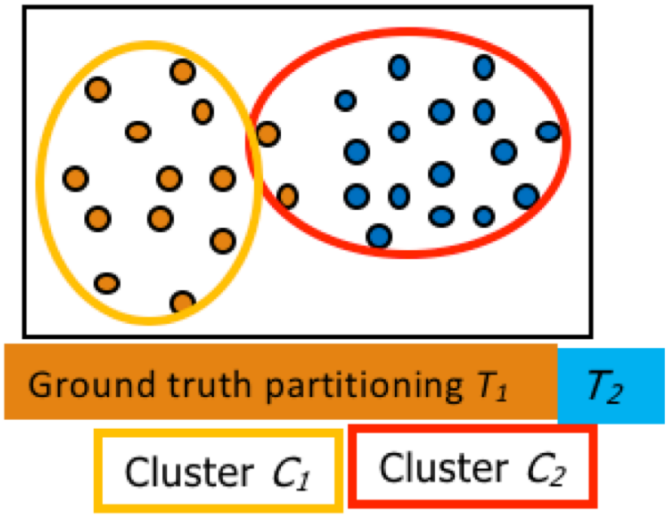
- No commonly recognized best suitable measure in practice
- **Three criteria**
  - **External**: Supervised, employ criteria not inherent to the dataset
    - Compare a clustering against prior or expert-specified knowledge (i.e., the ground truth) using certain clustering quality measure
  - **Internal**: Unsupervised, criteria derived from data itself
    - Evaluate the goodness of a clustering by considering how well the clusters are separated and how compact the clusters are, e.g., silhouette coefficient, squared-error
  - **Relative**: Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

Clustering Quality: External Methods

- Given the **ground truth** T, $Q(C, T)$ is the **quality measure** for a clustering C
- $Q(C, T)$ is good if it satisfies the following **four** essential criteria
  - **Cluster homogeneity:** The purer, the better
  - **Cluster completeness:** Assign objects belonging to the same category in the ground truth to the same cluster
  - **Rag bag better than alien:** Putting a heterogeneous object into a pure cluster should be penalized more than putting it into a *rag bag* (i.e., "miscellaneous" or "other" category)
  - **Small cluster preservation:** Splitting a small category into pieces is more harmful than splitting a large category into pieces

Commonly Used External Measures

- **Matching-based measures**
  - Purity, maximum matching, F-measure
- **Pairwise measures**
  - Four possibilities: True positive (TP), FN, FP, TN
  - Jaccard coefficient, Rand statistic, Fowlkes-Mallow measure
- **Entropy-Based Measures**
  - Conditional entropy
  - Normalized mutual information (NMI)



1. Matching-based

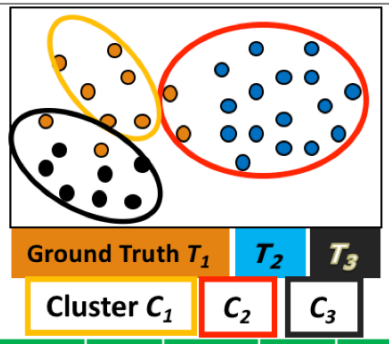
- **Purity**: Quantifies the extent that cluster C_i contains points only from one (ground truth) partition:

$$purity_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\}$$

  - Total purity of clustering C:

$$purity = \sum_{i=1}^{r} \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^{r} \max_{j=1}^{k} \{n_{ij}\}$$

  - Perfect clustering if purity = 1 and $r = k$ (the number of clusters obtained is the same as that in the ground truth)
  - Ex. 1 (green or orange): $purity_1$ = 30/50; $purity_2$ = 20/25; $purity_3$ = 25/25; $purity$ = (30 + 20 + 25)/100 = 0.75
  - ***Two clusters may share the same majority partition***



C\T	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

C\T	T_1	T_2	T_3	Sum
C_1	0	30	20	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	50	25	100

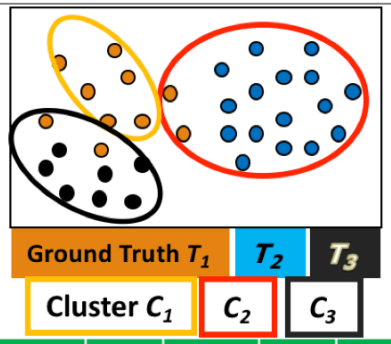
1. Matching-based (cont.)

- **Maximum matching**: ***Only one cluster can match one partition***
  - Match: Pairwise matching, weight $w(e_{ij}) = n_{ij}$
  - Maximum weight matching: Pair-wise

$$\mathrm{w(M)} = \sum_{e \in M} w(e)$$

$$match = \arg\max_{M} \{ \frac{w(M)}{n} \}$$

  - Ex2. (green) *match* = *purity* = 0.75;
    (orange) *match* = 0.65 > 0.6



C\T	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

C\T	T_1	T_2	T_3	Sum
C_1	0	30	20	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	50	25	100

1. Matching-based (cont.)

- **Precision**: The fraction of points in C_i from the majority partition T_{j_i} (i.e., the same as purity), where j_i is the partition that contains the maximum # of points from C_i

$$prec_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\} = \frac{n_{ij_i}}{n_i}$$

  - Ex. For the green table
    - $prec_1$ = 30/50; $prec_2$ = 20/25; $prec_3$ = 25/25

$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

1. Matching-based (cont.)

- **Recall**: The fraction of point in partition T_{j_i} shared in common with cluster C_i, where $m_{j_i} = |T_{j_i}|$

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$$

  - Ex. For the green table
    - $recall_1$ = 30/35; $recall_2$ = 20/40; $recall_3$ = 25/25

$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

1. Matching-based (cont.)

- **F-measure** for C_i: The harmonic means of $prec_i$ and $recall_i$:

$$F_i = \frac{2n_{ij_i}}{n_i + m_{j_i}}$$

- F-measure for clustering C: average of all clusters:

$$F = \frac{1}{r}\sum_{i=1}^{r} F_i$$

  - Ex. For the green table
    - F_1 = 60/85; F_2 = 40/65; F_3 = 1; F = 0.774

2. Pairwise

- Confusion matrix: **Four possibilities** based on the agreement between cluster label and partition label
  - *TP*: true positive—Two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same partition *T*, and they also in the same cluster *C*

$$TP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$$

where y_i: the true partition label, and $\hat{y}_i$: the cluster label for point $\boldsymbol{x}_i$

  - *FN*: false negative

$$FN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$$

  - *FP: false positive*

$$FP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$$

  - *TN*: true negative

$$TN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$$

2. Pairwise (cont.)

- Calculate the four measures:

$$N = \binom{n}{2}$$

Total # of pairs of points

$$TP = \sum_{i=1}^{r}\sum_{j=1}^{k}\binom{n_{ij}}{2} = \frac{1}{2}\left(\left(\sum_{i=1}^{r}\sum_{j=1}^{k} {n_{ij}}^2\right) - n\right)$$

$$FN = \sum_{j=1}^{k}\binom{m_j}{2} - TP$$

$$FP = \sum_{i=1}^{r}\binom{n_i}{2} - TP$$

$$TN = N - (TP + FN + FP) = \frac{1}{2}\left(n^2 - \sum_{i=1}^{r} {n_i}^2 - \sum_{j=1}^{k} {m_j}^2 + \sum_{i=1}^{r}\sum_{j=1}^{k} {n_{ij}}^2\right)$$

2. Pairwise (cont.)

- Jaccard coefficient: Fraction of true positive point pairs, but after ignoring the true negatives (thus asymmetric)
  - Jaccard = TP/(TP + FN + FP) [i.e., denominator ignores TN]
  - Perfect clustering: Jaccard = 1
- Rand Statistic:
  - Rand = (TP + TN)/*N_total*
  - Symmetric; perfect clustering: Rand = 1
- Fowlkes-Mallow Measure:
  - Geometric mean of precision and recall

$C\backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

$$FM = \sqrt{prec \times recall} = \frac{TP}{\sqrt{(TP+FN)(TP+FP)}}$$

- Using the above formulas, one can calculate all the measures for the green table (**leave as an exercise**)

3. Entropy-based

- **Entropy of clustering *C*:**

$$H(\mathcal{C}) = -\sum_{i=1}^{r} p_{C_i} \log p_{C_i} \qquad p_{C_i} = \frac{n_i}{n} \text{ (i.e., the probability of cluster } C_i)$$

- **Entropy of partitioning *T*:**

$$H(\mathcal{T}) = -\sum_{j=1}^{k} p_{T_i} \log p_{T_j}$$

- **Entropy of *T* with respect to cluster C_i:**

$$H(\mathcal{T}|C_i) = -\sum_{j=1}^{k} (\frac{n_{ij}}{n_i}) \log(\frac{n_{ij}}{n_i})$$

Conditional entropy

3. Entropy-based (cont.)

- **Conditional entropy of *T* with respect to clustering *C*:**

$$H(\mathcal{T}|\mathcal{C}) = -\sum_{i=1}^{r}(\frac{n_i}{n})H(\mathcal{T}|C_i) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log(\frac{p_{ij}}{p_{C_i}})$$

  - The more a cluster's members are split into different partitions, the higher the conditional entropy
  - For a perfect clustering, the conditional entropy value is o, where the worst possible conditional entropy value is *log k*

$$H(\mathcal{T}|\mathcal{C}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}(\log p_{ij} - \log p_{C_i}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log p_{ij} + \sum_{i=1}^{r}(\log p_{C_i}\sum_{j=1}^{k} p_{ij})$$

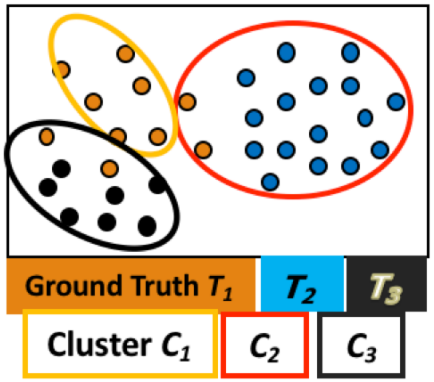
$$= -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log p_{ij} + \sum_{i=1}^{r}(p_{C_i}\log p_{C_i}) = H(\mathcal{C},\mathcal{T}) - H(\mathcal{C})$$

3. Entropy-based (cont.)

- **Mutual information**:
  - Quantifies the amount of shared info between the clustering *C* and partitioning *T*

$$I(C,T) = \sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log\left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}}\right)$$

  - Measures the dependency between the observed joint probability p_{ij} of *C* and *T*, and the expected joint probability $p_{Ci} \cdot p_{Tj}$ under the independence assumption
  - When *C* and *T* are independent, $p_{ij} = p_{Ci} \cdot p_{Tj}$, $I(C, T) = 0$. However, there is no upper bound on the mutual information



3. Entropy-based (cont.)

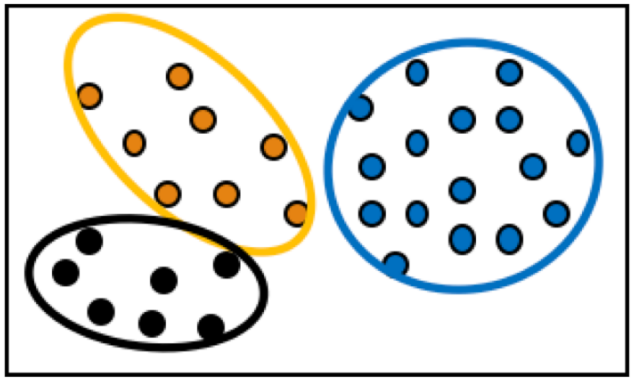
- **Normalized mutual information** (NMI)

$$NMI(\mathcal{C},\mathcal{T}) = \sqrt{\frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{C})} \cdot \frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{T})}} = \frac{I(\mathcal{C},\mathcal{T})}{\sqrt{H(\mathcal{C}) \cdot H(\mathcal{T})}}$$

  - Value range of NMI: [0,1]. Value close to 1 indicates a good clustering

Internal Methods: 1. BetaCV

- A trade-off in maximizing intra-cluster compactness and inter-cluster separation
- Given a clustering $C = \{C_1, \ldots, C_k\}$ with k clusters, cluster C_i containing $n_i = |C_i|$ points
  - Let $W(S, R)$ be sum of weights on all edges with one vertex in S and the other in R



1. BetaCV (cont.)

- The sum of all the intra-cluster weights over all clusters:

$$W_{in} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, C_i)$$

- The sum of all the inter-cluster weights:

$$W_{out} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1}\sum_{j>i} W(C_i, C_j)$$

- The number of distinct intra-cluster edges:

$$N_{in} = \sum_{i=1}^{k} \binom{n_i}{2}$$

- The number of distinct inter-cluster edges:

$$N_{out} = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j$$

1. BetaCV (cont.)

- **Beta-CV measure**:
  - The ratio of the mean intra-cluster distance/similarity to the mean inter-cluster distance/similarity

$$BetaCV = \frac{W_{in} / N_{in}}{W_{out} / N_{out}}$$

  - The smaller, the better the clustering, when the weight is distance
  - The bigger, the better the clustering, when the weight is similarity

2. Normalized Cut

- **Normalized cut**:

$$NC = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{vol(C_i)} = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{W(C_i, \mathrm{V})} = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{W(C_i, C_i) + W(C_i, \overline{C_i})} = \sum_{i=1}^{k} \frac{1}{\frac{W(C_i, C_i)}{W(C_i, \overline{C_i})} + 1}$$

where $vol(C_i) = W(C_i, V)$ is the volume of cluster C_i

– The higher normalized cut value, the better the clustering

3. Modularity

- **Modularity** (for graph clustering)
  - Modularity *Q* is defined as

$$Q = \sum_{i=1}^{k} \left(\frac{W(C_i, C_i)}{W(V,V)} - \left(\frac{W(C_i, V)}{W(V,V)} \right)^2 \right)$$

  where

$$W(V,V) = \sum_{i=1}^{k} W(C_i, V) = \sum_{i=1}^{k} W(C_i, C_i) + \sum_{i=1}^{k} W(C_i, \overline{C_i}) = 2(W_{in} + W_{out})$$

  - Modularity measures the difference between the observed and expected fraction of weights on edges within the clusters.
  - The smaller the value, the better the clustering—the intra-cluster distances are lower than expected

4. Silhouette Coefficient

- Check cluster cohesion and separation
  - For each point $\boldsymbol{x}_i$, its silhouette coefficient s_i is:

$$s_i = \frac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\}}$$

where $\mu_{in}(\mathbf{x}_i)$ is the mean distance from $\boldsymbol{x}_i$ to points in its own cluster

$\mu_{out}^{\min}(\mathbf{x}_i)$ is the mean distance from $\boldsymbol{x}_i$ to points in its closest cluster

  - Silhouette coefficient (*SC*) is the mean values of s_i across all the points:

$$SC = \frac{1}{n}\sum_{i=1}^{n} s_i$$

  - *SC* close to +1 implies good clustering
    - Points are close to their own clusters but far from other clusters

Relative Measure

- Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm
- **Silhouette coefficient** as a **relative measure**: Estimate the # of clusters in the data

$$SC_i = \frac{1}{n_i} \sum_{x_j \in C_i} s_j$$

Pick the *k* value that yields the best clustering, i.e., yielding high values for *SC* and SC_i ($1 \le i \le k$)

Reading Materials

- Defining and Evaluating Network Communities based on Ground-truth (ICDM'12)

https://cs.stanford.edu/people/jure/pubs/comscore-icdm12.pdf

References (1)

- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD'98
- M. R. Anderberg. Cluster Analysis for Applications. Academic Press, 1973.
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure, SIGMOD'99.
- Beil F., Ester M., Xu X.: "Frequent Term-Based Text Clustering", KDD'02
- M. M. Breunig, H.-P. Kriegel, R. Ng, J. Sander. LOF: Identifying Density-Based Local Outliers. SIGMOD 2000.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. KDD'96.
- M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. SSD'95.
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.

References (2)

- D. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2:139-172, 1987.
- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. VLDB'98.
- V. Ganti, J. Gehrke, R. Ramakrishan. CACTUS Clustering Categorical Data Using Summaries. KDD'99.
- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. In Proc. VLDB'98.
- S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. SIGMOD'98.
- S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. In ICDE'99, pp. 512-521, Sydney, Australia, March 1999.
- A. Hinneburg, D.l A. Keim: An Efficient Approach to Clustering in Large Multimedia Databases with Noise. KDD'98.

References (3)

- G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. COMPUTER, 32(8): 68-75, 1999.
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98.
- G. J. McLachlan and K.E. Bkasford. Mixture Models: Inference and Applications to Clustering. John Wiley and Sons, 1988.
- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.
- L. Parsons, E. Haque and H. Liu, Subspace Clustering for High Dimensional Data: A Review, SIGKDD Explorations, 6(1), June 2004
- E. Schikuta. Grid clustering: An efficient hierarchical clustering method for very large data sets. Proc. 1996 Int. Conf. on Pattern Recognition
- G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. VLDB'98.
- A. K. H. Tung, J. Han, L. V. S. Lakshmanan, and R. T. Ng. Constraint-Based Clustering in Large Databases, ICDT'01.

References (4)

- A. K. H. Tung, J. Hou, and J. Han. Spatial Clustering in the Presence of Obstacles, ICDE'01
- H. Wang, W. Wang, J. Yang, and P.S. Yu. Clustering by pattern similarity in large data sets, SIGMOD'02
- W. Wang, Yang, R. Muntz, STING: A Statistical Information grid Approach to Spatial Data Mining, VLDB'97
- T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH : An efficient data clustering method for very large databases. SIGMOD'96
- X. Yin, J. Han, and P. S. Yu, "LinkClus: Efficient Clustering via Heterogeneous Semantic Links", VLDB'06
- M. J. Zaki and W. Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014
- L. Hubert and P. Arabie. Comparing Partitions. Journal of Classification, 2:193–218, 1985

References (5)

- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988
- M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On Clustering Validation Techniques. Journal of Intelligent Info. Systems, 17(2-3):107–145, 2001
- J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- H. Xiong and Z. Li. Clustering Validation Measures. in (Chapter 23) C. Aggarwal and C. K. Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press, 2014