CLUSTER EVALUATION

OVERVIEW

- Motivation
- Evaluation type
  - Unsupervised
  - Supervised
- Unsupervised Evaluation
  - Cohesion
  - Separation
- Supervised Clustering /External Measure

MOTIVATION

MOTIVASI

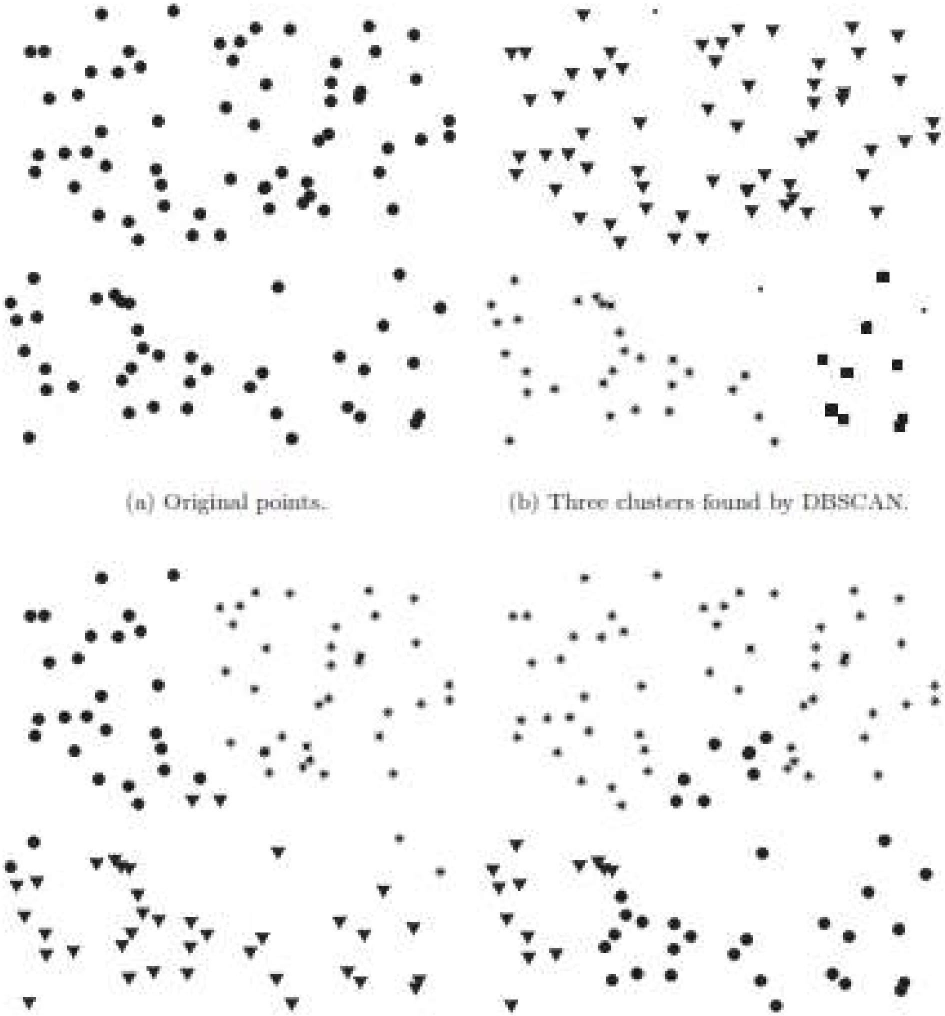
- Dalam supervised classification, evaluasi model klasifikasi yang dihasilkan adalah bagian integral dari proses pengembangan model klasifikasi, dan ada langkah-langkah evaluasi yang diterima dengan baik (misalnya, akurasi) dan prosedur (misalnya validasi silang).
- Akan tetapi, karena sifat dari clustering itu sendiri, cluster evaluation adalah sesuatu yang belum banyak dikembangkan atau banyak dipakai dalam analisa cluster

Motivasi

- Masing-masing algoritma mendefinisikan jenis clusternya sendiri, sehingga masing-masing clustering memerlukan evaluasi yg berbeda
  - For instance, K-means clusters might be evaluated in terms of the SSE, but for density-based clusters, which need not be globular, SSE would not work well at all.

MOTIVATION

- Cluster evaluation should be a part of any cluster analysis.
  - A key motivation is that almost every clustering algorithm will find clusters in a data set, even if that data set has no natural cluster structure.



(a) Original points.

(b) Three clusters found by DBSCAN.

(c) Three clusters found by K-means.

(d) Three clusters found by complete link.

OTHER MOTIVATIONS FOR CLUSTER VALIDATION

1. Menentukan kecenderungan pengelompokan satu set data, yaitu, membedakan apakah struktur non-acak benar-benar ada dalam data.
2. Menentukan jumlah cluster yang benar.
3. Mengevaluasi seberapa baik hasil analisis kluster sesuai dengan data tanpa referensi ke informasi eksternal.
4. Membandingkan hasil analisis kluster dengan hasil yang diketahui secara eksternal, seperti label kelas yang disediakan secara eksternal.
5. Membandingkan dua set cluster untuk menentukan mana yang lebih baik.

- Perhatikan bahwa item 1, 2, dan 3 tidak menggunakan informasi eksternal apa pun — supervised method— sedangkan item 4 membutuhkan informasi eksternal.
- Butir 5 dapat dilakukan dengan cara yang diawasi atau tidak diawasi.
- Perbedaan lebih lanjut dapat dibuat sehubungan dengan item 3, 4, dan 5: Apakah kita ingin mengevaluasi seluruh clustering atau hanya cluster individu?

TYPES OF EVALUATION

VARIUOUS TYPE OF CLUSTER VALIDITY

- Unsupervised
- Supervised
- Relative

UNSUPERVISED

- Measures the goodness of a clustering structure without respect to external information.
  - An example of this is the SSE.
- Unsupervised measures of cluster validity are often further divided into two classes:
  - measures of cluster cohesion (compactness, tightness), which determine how closely related the objects in a cluster are, and
  - Measures of cluster separation (isolation), which determine how distinct or well separated a cluster is from other clusters.
- Unsupervised measures are often called internal indices because they use only information present in the data set.

SUPERVISED

- Measures the extent to which the clustering structure discovered by a clustering algorithm matches some external structure.
- An example of a supervised index is entropy, which measures how well cluster labels match externally supplied class labels.
- Supervised measures are often called external indices because they use information not present in the data set.

RELATIVE

- Compares different clusterings or clusters.
- A relative cluster evaluation measure is a supervised or unsupervised evaluation measure that is used for the purpose of comparison.
- Thus, relative measures are not actually a separate type of cluster evaluation measure, but are instead a specific use of such measures.
- As an example, two K-means clusterings can be compared using either the SSE or entropy.

UNSUPERVISED

UNSUPERVISED CLUSTER EVALUATION USING COHESION AND SEPARATION

UNSUPERVISED CLUSTER EVALUATION USING THE PROXIMITY MATRIX.

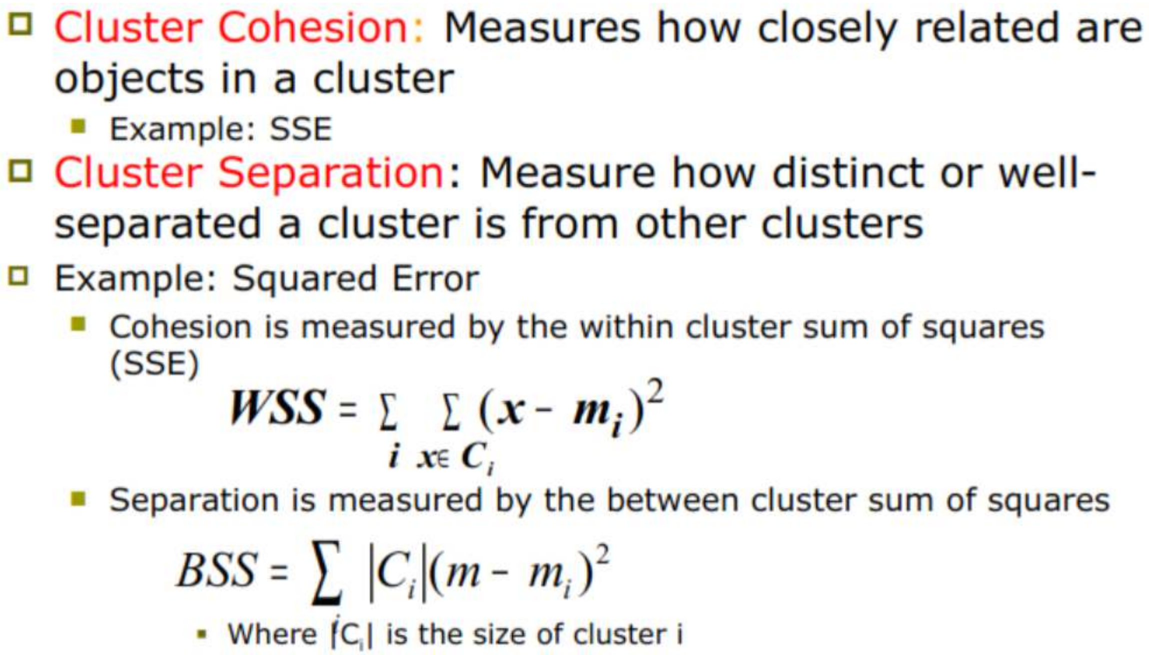
UNSUPERVISED CLUSTER EVALUATION USING COHESION AND SEPARATION

- Many internal measures of cluster validity for partitional clustering schemes are based on the notions of cohesion or separation.
- In general, we can consider expressing overall cluster validity for a set of K clusters as a weighted sum of the validity of individual clusters,

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

- The validity function can be cohesion, separation, or some combination of these quantities.
- The weights will vary depending on the cluster validity measure. In some cases, the weights are simply 1 or the size of the cluster, while in other cases they reflect a more complicated property, such as the square root of the cohesion.

PROTOTYPE-BASED VIEW OF 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 (SSE)

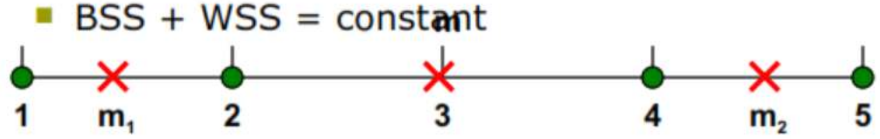
$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

  - Separation is measured by the between cluster sum of squares

$$BSS = \sum |C_i| (m - m_i)^2$$

    - Where $|C_i|$ is the size of cluster i

- Example: SSE
  - BSS + WSS = constant



K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

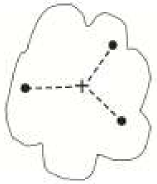
$$Total = 10 + 0 = 10$$

K=2 clusters:

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

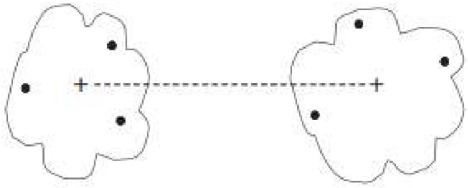
$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$



(a) Cohesion.

$$
\begin{aligned}
cohesion(C_i) &= \sum_{\mathbf{x} \in C_i} proximity(\mathbf{x}, \mathbf{c}_i) \\
separation(C_i, C_j) &= proximity(\mathbf{c}_i, \mathbf{c}_j) \\
separation(C_i) &= proximity(\mathbf{c}_i, \mathbf{c})
\end{aligned}
$$



(b) Separation.

- There are two measures for separation because the separation of cluster prototypes from an overall prototype is sometimes directly related to the separation of cluster prototypes from one another.

OVERALL MEASURES OF COHESION AND SEPARATION

- The previous definitions of cluster cohesion and separation gave us some simple and well defined measures of cluster validity that can be combined into an overall measure of cluster validity by using a weighted sum.
- However, we need to decide what weights to use.
- Not surprisingly, the weights used can vary widely, although typically they are some measure of cluster size.

- Note that any unsupervised measure of cluster validity potentially can be used as an objective function for a clustering algorithm and vice versa.
- The cluster evaluation measure I corresponds to traditional K-means and produces clusters that have good SSE values.
- The other measures produce clusters that are not as good with respect to SSE, but that are more optimal with respect to the specified cluster validity measure.

EVALUATING INDIVIDUAL CLUSTERS

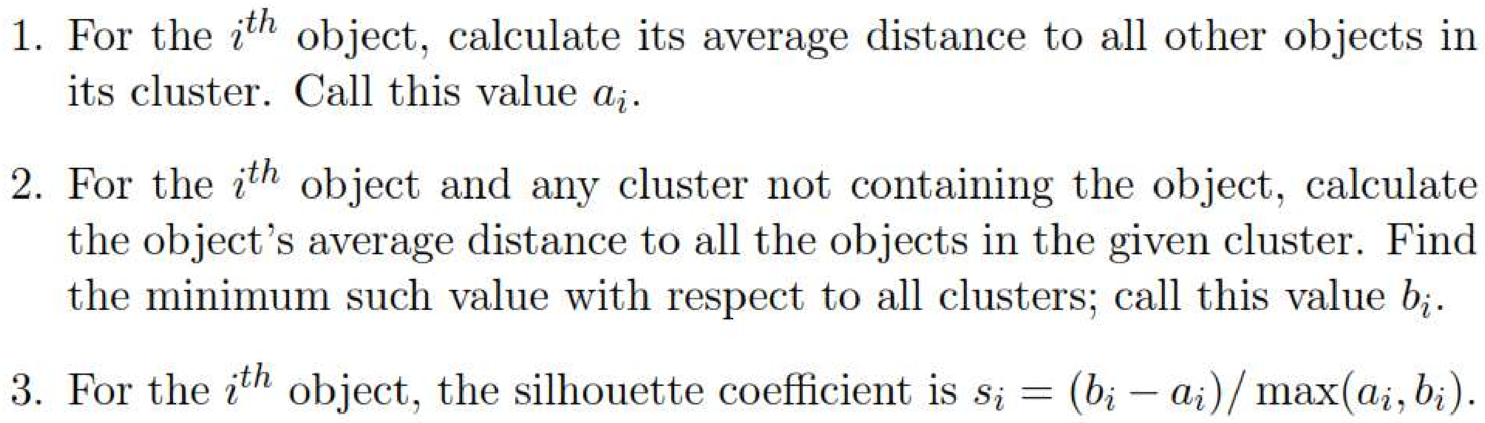
- Many of these measures of cluster validity also can be used to evaluate individual clusters.
- For example, we can rank individual clusters according to their specific value of cluster validity, i.e., cluster cohesion or separation.
- A cluster that has a high value of cohesion may be considered better than a cluster that has a lower value.
- This information often can be used to improve the quality of a clustering.
- If, for example, a cluster is not very cohesive, then we may want to split it into several sub clusters.
- On the other hand, if two clusters are relatively cohesive, but not well separated, we may want to merge them into a single cluster.

EVALUATING INDIVIDUAL CLUSTERS AND OBJECTS

- 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 cohesion and separation are near the "interior" of the cluster.
- Those objects for which the opposite is true are probably near the "edge" of the cluster.

THE SILHOUETTE COEfficIENT

- The popular method of silhouette coefficients combines both cohesion and separation.
- The following steps explain how to compute the silhouette coefficient for an individual point (we use distances, but an analogous approach can be used for similarities)

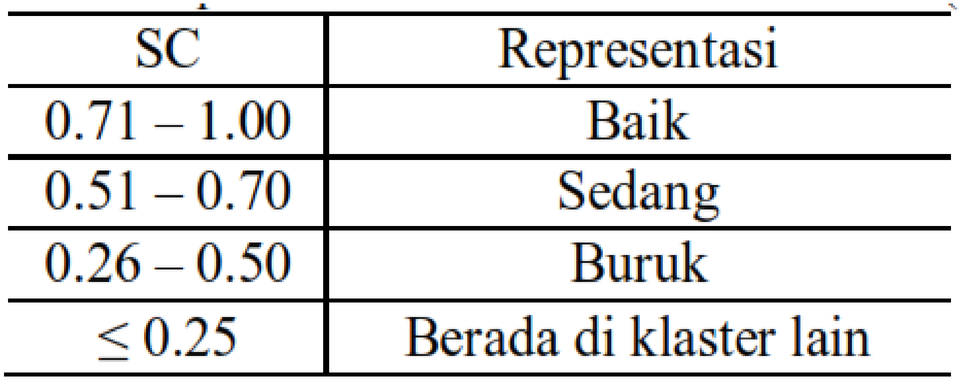


1. For the i^{th} object, calculate its average distance to all other objects in its cluster. Call this value a_i.

2. For the i^{th} object and any cluster not containing the object, calculate the 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.

3. For the i^{th} object, the silhouette coefficient is $s_i = (b_i - a_i)/\max(a_i, b_i)$.

The value of the silhouette coefficient can vary between -1 and 1. A negative value is undesirable because this corresponds to a case in which a_i, the average distance to points in the cluster, is greater than b_i, the minimum average distance to points in another cluster. We want the silhouette coefficient to be positive ($a_i < b_i$), and for a_i to be as close to 0 as possible, since the coefficient assumes its maximum value of 1 when $a_i = 0$.



SC	Representasi
0.71 – 1.00	Baik
0.51 – 0.70	Sedang
0.26 – 0.50	Buruk
≤ 0.25	Berada di klaster lain

- We can compute the average silhouette coefficient of a cluster by simply taking the average of the silhouette coefficients of points belonging to the cluster.
- An overall measure of the goodness of a clustering can be obtained by computing the average silhouette coefficient of all points.

2. UNSUPERVISED CLUSTER EVALUATION USING THE PROXIMITY MATRIX

- Measuring Cluster Validity via Correlation
- Judging a Clustering Visually by Its Similarity Matrix

MEASURING CLUSTER VALIDITY VIA CORRELATION

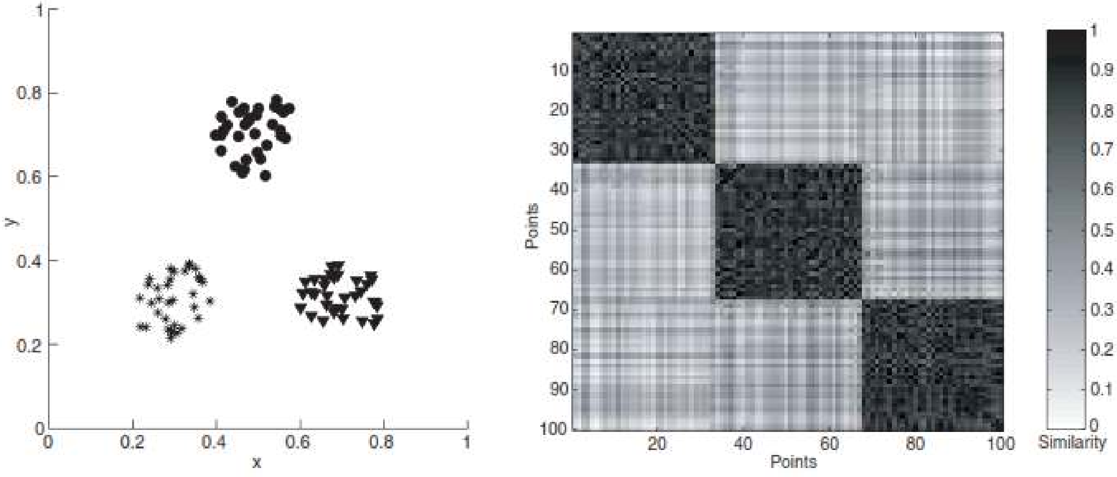
- If we are given the similarity matrix for a data set and the cluster labels from a cluster analysis of the data set, then we can evaluate the "goodness" of the clustering by looking at the correlation between the similarity matrix and an ideal version of the similarity matrix based on the cluster labels.
- (With minor changes, the following applies to proximity matrices, but for simplicity, we discuss only similarity matrices.)

- More specifically, an ideal cluster is one whose points have a similarity of 1 to all points in the cluster, and a similarity of 0 to all points in other clusters.
- Thus, if we sort the rows and columns of the similarity matrix so that all objects belonging to the same class are together, then an ideal similarity matrix has a block diagonal structure.
- In other words, the similarity is non-zero, i.e., 1, inside the blocks of the similarity matrix whose entries represent intra-cluster similarity, and 0 elsewhere.

- High correlation between the ideal and actual similarity matrices indicates that the points that belong to the same cluster are close to each other, while low correlation indicates the opposite.
- Since the actual and ideal similarity matrices are symmetric, the correlation is calculated only among the n(n-1)/2 entries below or above the diagonal of the matrices.
- Consequently, this is not a good measure for many density- or contiguity-based clusters, because they are not globular and may be closely intertwined with other clusters.

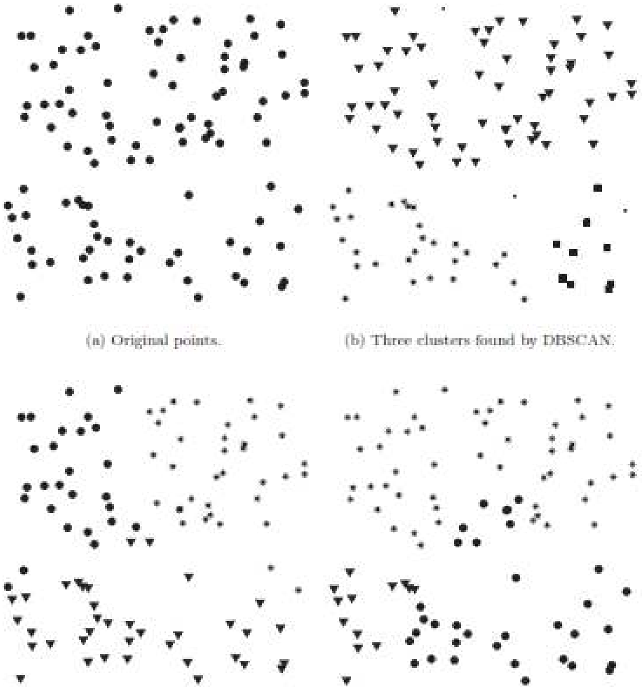
JUDGING A CLUSTERING VISUALLY BY ITS SIMILARITY MATRIX

- In theory, if we have well-separated clusters, then the similarity matrix should be roughly block-diagonal.



(a) Well-separated clusters.

(b) Similarity matrix sorted by K-means cluster labels.

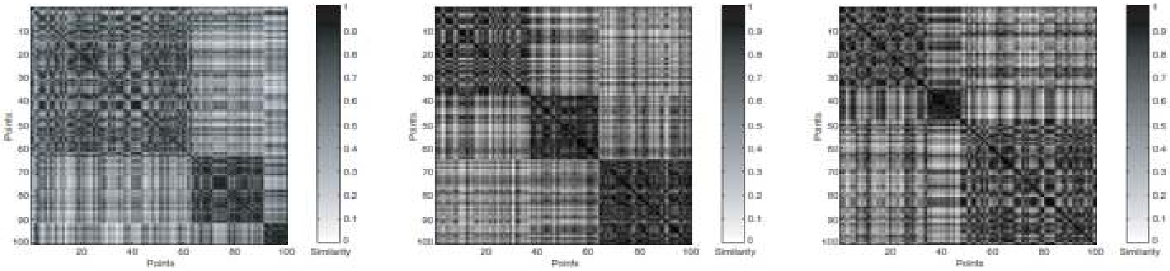


(a) Original points.

(b) Three clusters found by DBSCAN.

(c) Three clusters found by K-means.

(d) Three clusters found by complete link.



(a) Similarity matrix sorted by DBSCAN cluster labels.

(b) Similarity matrix sorted by K-means cluster labels.

(c) Similarity matrix sorted by complete link cluster labels.

- This approach may seem hopelessly expensive for large data sets, since the computation of the proximity matrix takes O(m^2) time, where m is the number of objects.
- Alternative: We can take a sample of data points from each cluster, compute the similarity between these points, and plot the result.
- It may be necessary to oversample small clusters and undersample large ones to obtain an adequate representation of all clusters.

DETERMINING THE CORRECT NUMBER OF CLUSTERS

- Various unsupervised cluster evaluation measures can be used to approximately determine the correct or natural number of clusters.
- Thus, we can try to find the natural number of clusters in a data set by looking for the number of clusters at which there is a knee, peak, or dip in the plot of the evaluation measure when it is plotted against the number of clusters.

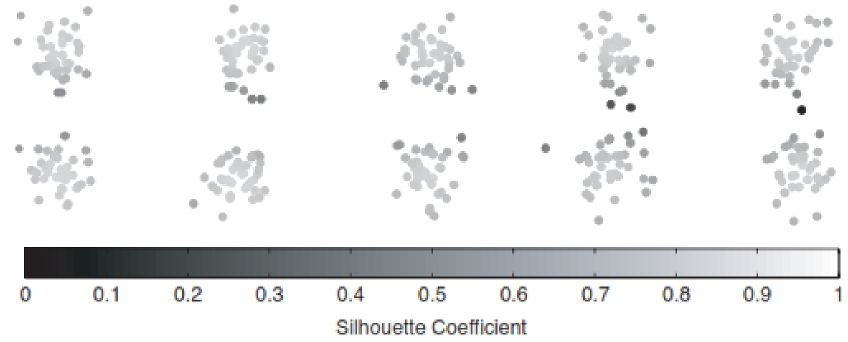


Figure 8.29. Silhouette coefficients for points in ten clusters.

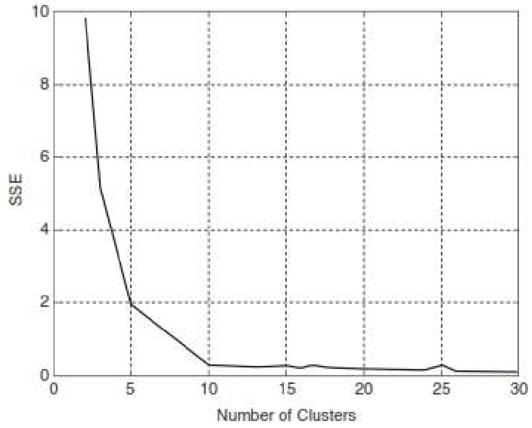


Figure 8.32. SSE versus number of clusters for the data of Figure 8.29.

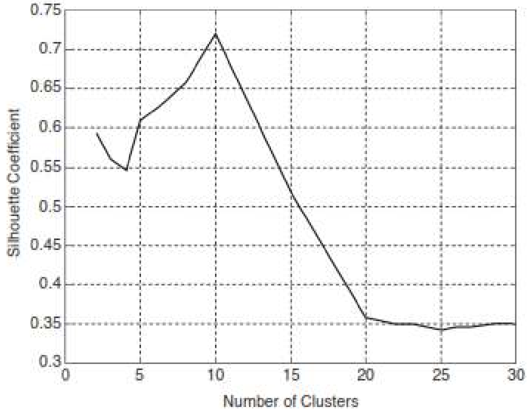


Figure 8.33. Average silhouette coefficient versus number of clusters for the data of Figure 8.29.

CLUSTERING TENDENCY

- One obvious way to determine if a data set has clusters is to try to cluster it.
- To address this issue, we could evaluate the resulting clusters and only claim that a data set has clusters if at least some of the clusters are of good quality.
- If the clusters are uniformly poor, then this may indeed indicate that there are no clusters in the data.

SUPERVISED METHOD

SUPERVISED MEASURES OF CLUSTER VALIDITY

- But why is this of interest? After all, if we have the class labels, then what is the point in performing a cluster analysis?
- Motivations for such an analysis are the comparison of clustering techniques with the "ground truth" or the evaluation of the extent to which a manual classification process can be automatically produced by cluster analysis.

CLASSIfiCATION-ORIENTED MEASURE FOR CLUSTER VALIDITY

- In the case of classification, we measure the degree to which predicted class labels correspond to actual class labels, but for the measures just mentioned, nothing fundamental is changed by using cluster labels instead of predicted class labels.
- Examples:
  - Entropy
  - Purity
  - Precision
  - Recall

ENTROPY

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

$p_{ij} = \frac{m_{ij}}{m_i}$ $\quad$ $e_i = -\sum_{j=1}^{L} p_{ij} log_2 p_{ij}$ $\quad$ $e = \sum_{i=1}^{K} \frac{m_i}{m} e_i$

Keterangan:

- p_{ij} : the probability that a member of cluster i belongs to class j
- m_{ij} : the number of objects of class j in cluster i
- m_i : the number of objects in cluster i
- e_i : entropy of cluster i
- e : the total entropy
- L is the number class; K is the number of cluster

PURITY

Another measure of the extent to which a cluster contains objects of a single class.

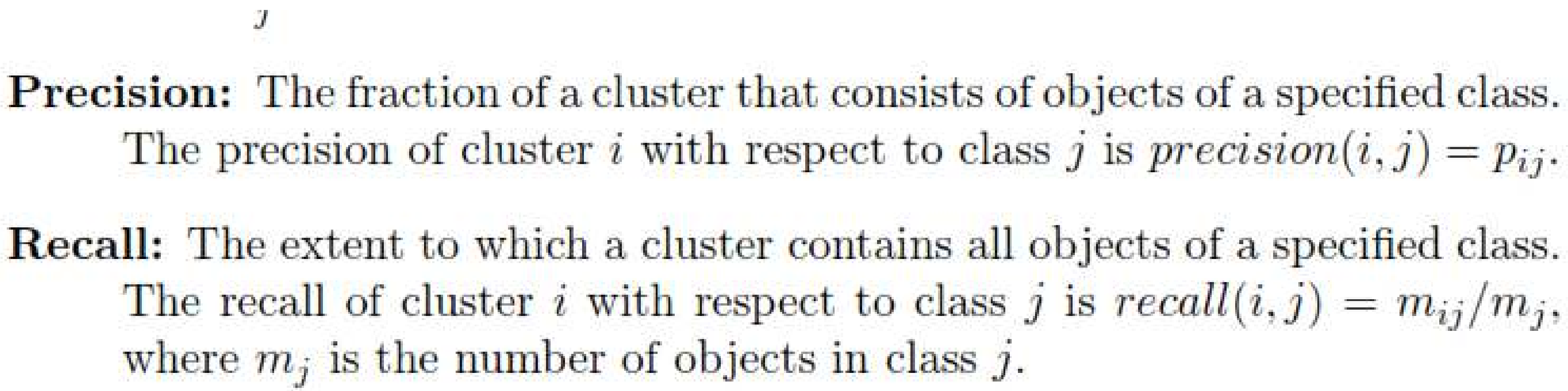
$p_i = \max_j p_{ij}$ $\text{purity} = \sum_{i=1}^{K} \frac{m_i}{m} p_i$

Keterangan:

p_i : purity of cluster I

purity : the total purity

PRECISION AND RECALL



j

Precision: The fraction of a cluster that consists of objects of a specified class. The precision of cluster i with respect to class j is $precision(i, j) = p_{ij}$.

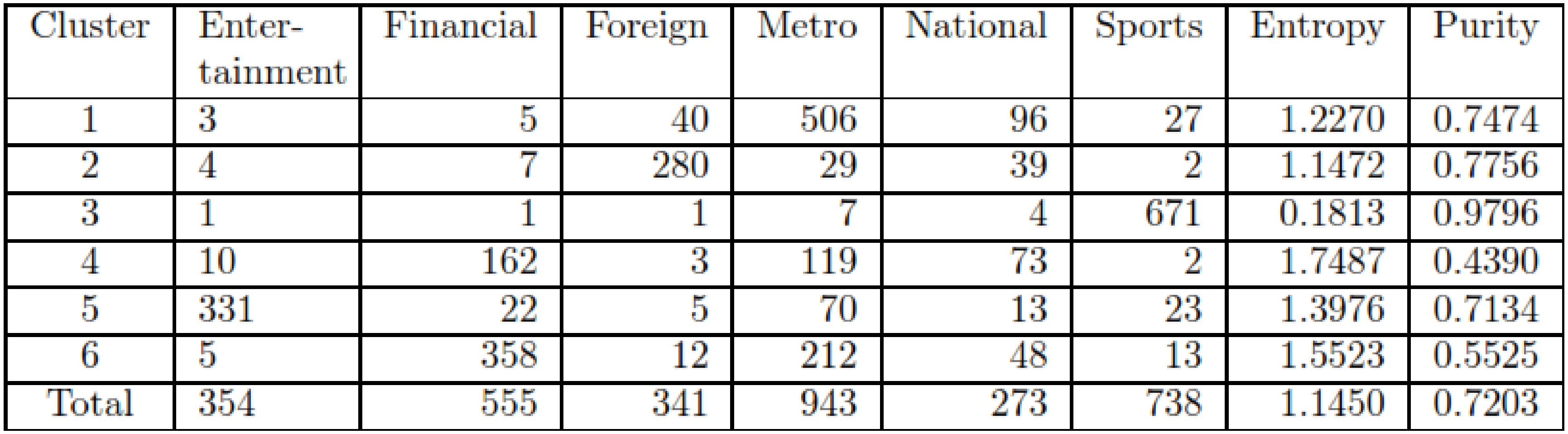
Recall: The extent to which a cluster contains all objects of a specified class. The recall of cluster i with respect to class j is $recall(i, j) = m_{ij}/m_j$, where m_j is the number of objects in class j.

- A true positive (TP) decision assigns two similar documents to the same cluster, a true negative (TN) decision assigns two dissimilar documents to different clusters. There are two types of errors we can commit.
- A (FP) decision assigns two dissimilar documents to the same cluster. A (FN) decision assigns two similar documents to different clusters.
- The *Rand index* ( ) measures the percentage of decisions that are correct.
-

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Table 8.9. K-means clustering results for the *LA Times* 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

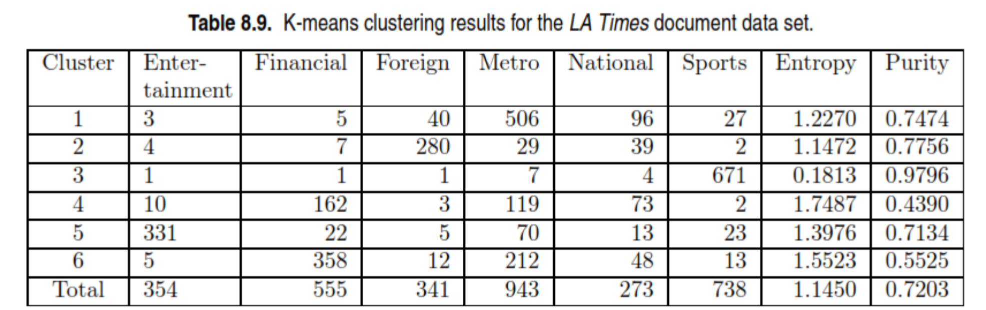


Table 8.9. K-means clustering results for the *LA Times* document data set.

Cluster	Enter-tainment	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

- P1, metro= 506/677
- P1,entertainment=3/677
- P1,financial=/677
- P1,foreign=40/677
- P1,National=96/677
- P1Sports=27/677
- E1=

- Example: consider cluster 1 and the Metro class of Table 8.9.
- The precision is 506/677 = 0.75
- The recall is 506/943 = 0.26