CS522 Advanced Database Systems
Clustering: Cluster Evaluation

Chengyu Sun
California State University, Los Angeles

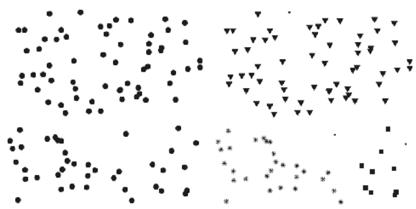
Cluster Evaluation

- A.K.A. *Cluster Validation*
- Unsupervised
  - Using no external information other than the data itself
- Supervised
  - With external information such as given class labels

Reasons Not To Evaluate

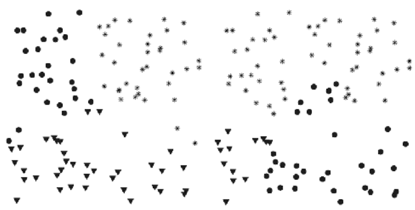
- Clustering is often used as part of exploratory data analysis
- Clustering is often used as part of other algorithms
- Clustering algorithms, in some sense, define their own types of clusters

Reasons To Evaluate ...



(a) Original points. (b) Three clusters found by DBSCAN.

... Reasons To Evaluate



(c) Three clusters found by K-means. (d) Three clusters found by complete link.

Questions To Be Answered

- Do clusters actually exist?
- How many clusters are there?
- How good is a cluster/clustering?

Clustering Tendency

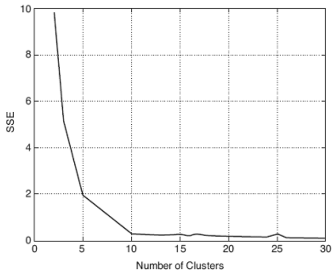
- Whether clusters exist in the first place
- Determine clustering tendency
  - Cluster first, then evaluate the quality of the clustering
    - Need to try several different types of clustering algorithms
  - Statistical tests for spatial randomness

Hopkins Statistic

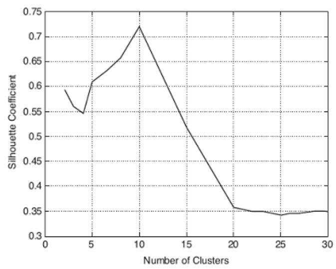
- Generate `p` random points in the data space
  - u_i: distance of a randomly generated point to its nearest neighbor in the original dataset
- Select `p` random points from the original dataset
  - w_i: distance of a randomly selected point to it nearest neighbor in the original dataset
- *Interpretation of Hopkins Statistic??*

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

Determine The Correct Number of Clusters ...



... Determine The Correct Number of Clusters



Quality (Validity) of Clusters

- Cohesion
  - Compactness of a cluster
- Separation

Validity of Prototype-based Clusters

$$cohesion(C_i) = \sum_{\mathbf{x} \in C_i} dist(\mathbf{x}, \mathbf{c_i})$$

$$separation(C_i, C_j) = dist(\mathbf{c}_i, \mathbf{c}_j)$$

$$separation(C_i) = dist(\mathbf{c}_i, \mathbf{c})$$

Validity of Graph-based Clusters

$$cohesion(C_i) = \sum_{\substack{\mathbf{x}\in C_i \\ \mathbf{y}\in C_i}} dist(\mathbf{x},\mathbf{y})$$

$$separation(C_i, C_j) = \sum_{\substack{\mathbf{x}\in C_i \\ \mathbf{y}\in C_j}} dist(\mathbf{x},\mathbf{y})$$

Validity of A Clustering

$$validity(C) = \sum_{i=1}^{k} w_i \times validity(C_i)$$

Cluster Weights

Validity Measures	Weights
$\sum_{\substack{\mathbf{x}\in C_i \\ \mathbf{y}\in C_i}} dist(\mathbf{x},\mathbf{y})$	$1/\lvert C_i \rvert$
$\sum_{\mathbf{x}\in C_i} dist(\mathbf{x},\mathbf{c_i})$	1
$dist(\mathbf{c}_i,\mathbf{c})$	$\lvert C_i \rvert$

Silhouette Coefficient

- For the `i`th object in a cluster
  - a_i: average distance to all other objects in the cluster
  - b_i: minimum of the average distance to the objects in a cluster that does not contain this object

$$s_i = (b_i - a_i)/\max(a_i, b_i)$$

About Silhouette Coefficient

- *Range of s_i??*
- *What is a "good" value of s_i??*
- Quality of an object: s_i
- Quality of a cluster/clustering: average s_i

Silhouette Coefficient Example

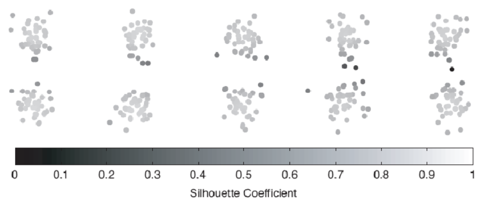
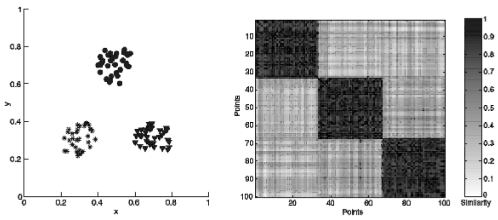


Figure 8.29. Silhouette coefficients for points in ten clusters.

Similarity Matrix

- Sort the objects by cluster label
- Similarity Matrix **M**
  - M(i,j) = similarity($\mathbf{x_i}$,$\mathbf{x_j}$), 0≤M(i,j)≤1

Visualizing Clustering Results Using Similarity Matrix



(a) Well-separated clusters.

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

Supervised Measures of Cluster Validity

- Classification-oriented measures
  - Evaluate the extent to which a cluster contains the objects of a single class
- Similarity-oriented measures
  - Evaluate the extent to which two objects of the same class (or cluster) belong to the same cluster (or class)

Classification-Oriented Measures

- Entropy
- Purity
- Precision, recall, F-measure

Similarity-Oriented Measures – Contingency Table

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

f – the number of *pairs* of objects

Example

- Classes: {p_1,p_2}, {p_3,p_4,p_5}
- Clusters: {p_1,p_2,p_3}, {p_4,p_5}

Similarity Measures

Rand Statistic: $$R = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Jaccard Coefficient: $$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$