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 ...

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 its nearest neighbor in the original dataset
- Interpretation of Hopkins Statistic?

\[
H = \frac{\sum w_i}{\sum u_i + \sum w_i}
\]

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

\[
cohesion(C_i) = \sum_{x \in C_i} \text{dist}(x, c_i)
\]
\[
separation(C_i, C_j) = \text{dist}(c_i, c_j)
\]
\[
separation(C_i) = \text{dist}(c_i, c)
\]
Validity of Graph-based Clusters

\[
\text{cohesion}(C_i) = \sum_{\substack{x \in C_i \setminus \{c_i\}, \; y \in C_i}} \text{dist}(x, y)
\]

\[
\text{separation}(C_i, C_j) = \sum_{\substack{x \in C_i \setminus \{c_i\}, \; y \in C_j}} \text{dist}(x, y)
\]

Validity of A Clustering

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

Cluster Weights

<table>
<thead>
<tr>
<th>Validity Measures</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>[\sum_{\substack{x \in C_i \setminus {c_i}, ; y \in C_i}} \text{dist}(x, y)]</td>
<td>[1/</td>
</tr>
<tr>
<td>[\sum_{\substack{x \in C_i \setminus {c_i}}} \text{dist}(x, c_i)]</td>
<td>[1]</td>
</tr>
<tr>
<td>[\text{dist}(c_i, c)]</td>
<td>[</td>
</tr>
</tbody>
</table>

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 $M$
  - $M(i,j) = \text{similarity}(x_i, x_j)$, $0 \leq M(i,j) \leq 1$

Visualizing Clustering Results Using Similarity Matrix

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

Example

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

Similarity-Oriented Measures – Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Same cluster</th>
<th>Different cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same class</td>
<td>$f_{11}$</td>
<td>$f_{10}$</td>
</tr>
<tr>
<td>Different class</td>
<td>$f_{01}$</td>
<td>$f_{00}$</td>
</tr>
</tbody>
</table>

$f$ – the number of pairs of objects
Similarity Measures

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

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