What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

Intra-cluster distances are minimized
Inter-cluster distances are maximized

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Applications of Cluster Analysis

- Understanding
  - Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

- Summarization
  - Reduce the size of large data sets

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Notion of a Cluster can be Ambiguous

- How many clusters?
- Six Clusters
- Two Clusters
- Four Clusters

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Types of Clusterings

- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
- Partitional Clustering
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Partitional Clustering

Original Points
A Partitional Clustering

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Hierarchical Clustering

K-means Clustering

• Partitional clustering approach
• Each cluster is associated with a centroid (center point)
• Each point is assigned to the cluster with the closest centroid
• Number of clusters, K, must be specified
• The basic algorithm is very simple

1: Select K points as the initial centroids.
2: assigned
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: until The centroids don’t change

Two different K-means Clusterings

Importance of Choosing Initial Centroids
Clustering fundamentals

To get SSE, we square these errors and sum them.

\[ \text{SSE} = \sum_{i} \sum_{c} \text{dist}(m_c, x_i) \]

- \( m_c \) is a data point in cluster \( c \) and \( m_i \) is the representative point for cluster \( c \).
- \( x_i \) is a data point in cluster \( c \) and \( m_i \) is the representative point for cluster \( c \).
- One easy way to reduce SSE is to increase \( K \), the number of clusters. A good clustering with smaller \( K \) can have a lower SSE than a poor clustering with higher \( K \).

Limitations of K-means

- K-means has problems when clusters are of differing:
  - Sizes
  - Densities
  - Non-globular shapes

- K-means has problems when the data contains outliers.

Importance of Choosing Initial Centroids

Most common measure is Sum of Squared Error (SSE)

- For each point, the error is the distance to the nearest cluster.
- To get SSE, we square these errors and sum them.

Evaluating K-means Clusters

- For each point, the error is the distance to the nearest cluster.
- To get SSE, we square these errors and sum them.

Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on your side.
- Sample and use hierarchical clustering to determine initial centroids.
- Select more than \( K \) initial centroids and then select among these initial centroids.
- Select most widely separated.
- Postprocessing
  - Bisecting K-means
  - Not as susceptible to initialization issues.

Pre-processing and Post-processing

- Pre-processing
  - Normalize the data
  - Eliminate outliers
- Post-processing
  - Eliminate small clusters that may represent outliers.
  - Split 'loose' clusters, i.e., clusters with relatively high SSE.
  - Merge clusters that are 'close' and that have relatively low SSE.

Database and Data Mining Group
Limitations of K-means: Differing Sizes

Original Points
K-means (3 Clusters)

Limitations of K-means: Differing Density

Original Points
K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

Original Points
K-means (2 Clusters)

Overcoming K-means Limitations

One solution is to use many clusters. Find parts of clusters, but need to put together.

Overcoming K-means Limitations

Original Points
K-means Clusters
Hierarchical Clustering
- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree-like diagram that records the sequences of merges or splits

Strengths of Hierarchical Clustering
- Do not have to assume any particular number of clusters
- Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level
- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, ...)

Hierarchical Clustering
- Two main types of hierarchical clustering
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)
- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

Agglomerative Clustering Algorithm
- More popular hierarchical clustering technique
- Basic algorithm is straightforward
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. Repeat
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. Until only a single cluster remains
- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms

How to Define Inter-Cluster Similarity
- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error
**DBSCAN**

- **DBSCAN** is a density-based algorithm.
  - Density = number of points within a specified radius (Eps)
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
  - These are points that are at the interior of a cluster
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point.

---

**When DBSCAN Works Well**

- Resistant to Noise
- Can handle clusters of different shapes and sizes

---

**When DBSCAN Does NOT Work Well**

- Varying densities
- High-dimensional data

---

**Measures of Cluster Validity**

- The validation of clustering structures is the most difficult task
- To evaluate the "goodness" of the resulting clusters, some numerical measures can be exploited
- Numerical measures are classified into two main classes
  - **External Index**: Used to measure the extent to which cluster labels match externally supplied class labels.
    - e.g., entropy, purity
  - **Internal Index**: Used to measure the goodness of a clustering structure without respect to external information.
    - e.g., Sum of Squared Error (SSE), cluster cohesion, cluster separation, Rand-Index, adjusted rand-index
### External Measures of Cluster Validity: Entropy and Purity

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Entertainment</th>
<th>Financial</th>
<th>Foreign</th>
<th>Movie</th>
<th>National</th>
<th>Sporty</th>
<th>Entropy</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>7</td>
<td>205</td>
<td>894</td>
<td>23</td>
<td>27</td>
<td>3.2977</td>
<td>0.7671</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
<td>29</td>
<td>48</td>
<td>27</td>
<td>2</td>
<td>3.677</td>
<td>0.7504</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>6</td>
<td>17</td>
<td>15</td>
<td>2.9512</td>
<td>0.9704</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>302</td>
<td>5</td>
<td>107</td>
<td>25</td>
<td>2</td>
<td>3.6467</td>
<td>0.4546</td>
</tr>
<tr>
<td>5</td>
<td>311</td>
<td>22</td>
<td>5</td>
<td>78</td>
<td>13</td>
<td>25</td>
<td>1.8879</td>
<td>0.7324</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>104</td>
<td>12</td>
<td>152</td>
<td>48</td>
<td>22</td>
<td>3.1022</td>
<td>0.5505</td>
</tr>
<tr>
<td>Total</td>
<td>354</td>
<td>555</td>
<td>341</td>
<td>443</td>
<td>738</td>
<td>708</td>
<td>3.154</td>
<td>0.7401</td>
</tr>
</tbody>
</table>

Entropy: For each cluster, the class distribution of the data is calculated first, i.e., for cluster \( j \) we compute \( p_{ij} \), the probability that a member of cluster \( j \) belongs to class \( i \) as follows:

\[
p_{ij} = \frac{n_{ij}}{n_j},
\]

where \( n_{ij} \) is the number of values in class \( i \) and \( n_j \) is the number of values in cluster \( j \). Then using this class distribution, the entropy of each cluster \( j \) is calculated using the standard formula:

\[
S_j = -\sum_i p_{ij} \log_2 p_{ij},
\]

where \( i \) is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropy of each cluster weighted by the size of each cluster: i.e.,

\[
S = \sum_j \frac{S_j}{n_j},
\]

where \( n_j \) is the size of cluster \( j \), \( n \) is the number of clusters, and \( n \) is the total number of data points.

Purity: Using the terminology derived for entropy, the purity of cluster \( j \), is given by purity \( = \max_{i \neq j} p_{ij} \) and the overall purity of a clustering by purity \( = \frac{\sum_i n_i p_{ij}}{n} \).

### 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 weights of all links within a cluster.
- Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.

### 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