Data Mining Techniques: Cluster Analysis

Mirek Riedewald

Many slides based on presentations by Han/Kamber, Tan/Steinbach/Kumar, and Andrew Moore

Cluster Analysis Overview

- Introduction
- Foundations: Measuring Distance (Similarity)
- Partitioning Methods: K-Means
- Hierarchical Methods
- Density-Based Methods
- Clustering High-Dimensional Data
- Cluster Evaluation

What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Unsupervised learning: usually no training set with known “classes”
- Typical applications
  - As a stand-alone tool to get insight into data properties
  - As a preprocessing step for other algorithms

Rich Applications, Multidisciplinary Efforts

- Pattern Recognition
- Spatial Data Analysis
- Image Processing
- Data Reduction
- Economic Science
  - Market research
- WWW
  - Document classification
  - Weblogs: discover groups of similar access patterns

Examples of Clustering Applications

- **Marketing**: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Land use**: Identification of areas of similar land use in an earth observation database
- **Insurance**: Identifying groups of motor insurance policy holders with a high average claim cost
- **City-planning**: Identifying groups of houses according to their house type, value, and geographical location
- **Earthquake studies**: Observed earth quake epicenters should be clustered along continent faults

What is Cluster Analysis?

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

Examples of Clustering Applications

- **Marketing**: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Land use**: Identification of areas of similar land use in an earth observation database
- **Insurance**: Identifying groups of motor insurance policy holders with a high average claim cost
- **City-planning**: Identifying groups of houses according to their house type, value, and geographical location
- **Earthquake studies**: Observed earthquake epicenters should be clustered along continent faults
Quality: What Is Good Clustering?

- Cluster membership ≈ objects in same class
- High intra-class similarity, low inter-class similarity
  - Choice of similarity measure is important
- Ability to discover some or all of the hidden patterns
  - Difficult to measure without ground truth

Notion of a Cluster can be Ambiguous

Distinctions Between Sets of Clusters

- Exclusive versus non-exclusive
  - Non-exclusive clustering: points may belong to multiple clusters
- Fuzzy versus non-fuzzy
  - Fuzzy clustering: a point belongs to every cluster with some weight between 0 and 1
    - Weights must sum to 1
- Partial versus complete
  - Cluster some or all of the data
- Heterogeneous versus homogeneous
  - Clusters of widely different sizes, shapes, densities

Cluster Analysis Overview

- Introduction
- Foundations: Measuring Distance (Similarity)
- Partitioning Methods: K-Means
- Hierarchical Methods
- Density-Based Methods
- Clustering High-Dimensional Data
- Cluster Evaluation

Distance

- Clustering is inherently connected to question of (dis-)similarity of objects
- How can we define similarity between objects?

Similarity Between Objects

- Usually measured by some notion of distance
- Popular choice: Minkowski distance
  \[ \text{dist}(x_i, x_j) = \left( \sum_{d=1}^{D} |x_i(d) - x_j(d)|^q \right)^{1/q} \]
  - \( q \) is a positive integer
- \( q = 1 \): Manhattan distance
  \[ \text{dist}(x_i, x_j) = |x_1(i) - x_1(j)| + |x_2(i) - x_2(j)| + \cdots + |x_d(i) - x_d(j)| \]
- \( q = 2 \): Euclidean distance:
  \[ \text{dist}(x_i, x_j) = \sqrt{\sum_{d=1}^{D} |x_i(d) - x_j(d)|^2} \]
**Metrics**

- Properties of a metric
  - \( d(i,j) \geq 0 \)
  - \( d(i,j) = 0 \) if and only if \( i=j \)
  - \( d(i,j) = d(j,i) \)
  - \( d(i,j) \leq d(i,k) + d(k,j) \)
- Examples: Euclidean distance, Manhattan distance
- Many other non-metric similarity measures exist
- After selecting the distance function, is it now clear how to compute similarity between objects?

**Challenges**

- How to compute a distance for categorical attributes
- An attribute with a large domain often dominates the overall distance
  - Weight and scale the attributes like for k-NN
- Curse of dimensionality

---

**Curse of Dimensionality**

- Best solution: remove any attribute that is known to be very noisy or not interesting
- Try different subsets of the attributes and determine where good clusters are found

**Nominal Attributes**

- Method 1: work with original values
  - Difference = 0 if same value, difference = 1 otherwise
- Method 2: transform to binary attributes
  - New binary attribute for each domain value
  - Encode specific domain value by setting corresponding binary attribute to 1 and all others to 0

**Ordinal Attributes**

- Method 1: treat as nominal
  - Problem: loses ordering information
- Method 2: map to [0,1]
  - Problem: To which values should the original values be mapped?
  - Default: equi-distant mapping to [0,1]

**Scaling and Transforming Attributes**

- Sometimes it might be necessary to transform numerical attributes to [0,1] or use another normalizing transformation, maybe even non-linear (e.g., logarithm)
- Might need to weight attributes differently
- Often requires expert knowledge or trial-and-error
Other Similarity Measures

- Special distance or similarity measures for many applications
  – Might be a non-metric function
- Information retrieval
  – Document similarity based on keywords
- Bioinformatics
  – Gene features in micro-arrays

Calculating Cluster Distances

- Single link = smallest distance between an element in one cluster and an element in the other: dist\( (K_i, K_j) = \min_{x_{ip}, x_{jq}} \) x
- Complete link = largest distance between an element in one cluster and an element in the other: dist\( (K_i, K_j) = \max_{x_{ip}, x_{jq}} \) x
- Average distance between an element in one cluster and an element in the other: dist\( (K_i, K_j) = \frac{\text{avg}(x_{ip}, x_{jq})}{2} \)
- Distance between cluster centroids: dist\( (K_i, K_j) = d(m_i, m_j) \)
- Distance between cluster medoids: dist\( (K_i, K_j) = \text{dist}(x_{mi}, x_{mj}) \)
  – Medoid: one chosen, centrally located object in the cluster

Cluster Centroid, Radius, and Diameter

- Centroid: the “middle” of a cluster \( C \)
  \[ m = \frac{1}{|C|} \sum_{x \in C} x \]
- Radius: square root of average distance from any point of the cluster to its centroid
  \[ R = \sqrt{\frac{\sum_{x \in C} (x - m)^2}{|C|}} \]
- Diameter: square root of average mean squared distance between all pairs of points in the cluster
  \[ D = \sqrt{\frac{\sum_{x, y \in C} (x - y)^2}{|C| - 1}} \]

Cluster Analysis Overview

- Introduction
- Foundations: Measuring Distance (Similarity)
- Partitioning Methods: K-Means
- Hierarchical Methods
- Density-Based Methods
- Clustering High-Dimensional Data
- Cluster Evaluation

Partitioning Algorithms: Basic Concept

- Construct a partition of a database \( D \) of \( n \) objects into a set of \( K \) clusters, s.t. sum of squared distances to cluster “representative” \( m \) is minimized
  \[ \sum_{i=1}^{K} \sum_{x \in C_i} (m_i - x)^2 \]
- Given a \( K \), find partition of \( K \) clusters that optimizes the chosen partitioning criterion
  – Globally optimal: enumerate all partitions
  – Heuristic methods
    - K-means ('67): each cluster represented by its centroid
    - K medoids ('87): each cluster represented by one of the objects in the cluster

K-means Clustering

- Each cluster is associated with a centroid
- Each object is assigned to the cluster with the closest centroid
1. Given \( K \), select \( K \) random objects as initial centroids
2. Repeat until centroids do not change
   1. Form \( K \) clusters by assigning every object to its nearest centroid
   2. Recompute centroid of each cluster
Distance usually measured by Euclidean
What is it trying to optimize?
So each time the configuration changes it must go to a configuration it
If the configuration changes on an iteration, it must have improved SSE
Easy way to reduce SSE: increase K
Most common measure: Sum of Squared Error (SSE)

SSE = \sum_{i=1}^{K} \sum_{x \in C_i} \text{dist}^2(m_i, x)

m_i = \text{centroid of cluster } C_i

Given two clusterings, choose the one with the smallest error
Easy way to reduce SSE: increase K
In practice, large K not interesting

Will it always terminate?
In practice, large K not interesting
Comparably fast algorithm: O(n * K * I * d)

• Initial centroids often chosen randomly
  Clusters produced vary from one run to another
• Distance usually measured by Euclidean
distance, cosine similarity, correlation, etc.
• Comparably fast algorithm: O(n * K * I * d)
  – n = number of objects
  – I = number of iterations
  – d = number of attributes

K-means Questions
• What is it trying to optimize?
• Will it always terminate?
• Will it find an optimal clustering?
• How should we start it?
• How could we automatically choose the number of centers?

...we’ll deal with these questions next

K-means Clustering Details

Evaluating K-means Clusters
• Most common measure: Sum of Squared Error (SSE)
  – For each point, the error is the distance to the nearest centroid
  \[ \text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \text{dist}^2(m_i, x) \]
  – \( m_i \) = centroid of cluster \( C_i \)
• Given two clusterings, choose the one with the smallest error
• Easy way to reduce SSE: increase K
  – In practice, large K not interesting

K-means Convergence
• (1) Assign each \( x \) to its nearest center (minimizes SSE for fixed centers)
• (2) Choose centroid of all points in the same cluster as cluster center (minimizes SSE for fixed clusters)
• Cycle through steps (1) and (2) = K-means algorithm
• Algorithm terminates when neither (1) nor (2) results in change of configuration
  – Finite number of ways of partitioning \( n \) records into \( K \) groups
  – If the configuration changes on an iteration, it must have improved SSE
  – So each time the configuration changes it must go to a configuration it has never been to before
  – So if it tried to go on forever, it would eventually run out of configurations
**Will it Find the Optimal Clustering?**

**Importance of Initial Centroids**

- Original Points

**Will It Find The Optimal Clustering Now?**

**Importance of Initial Centroids**

- Iteration 5

**Problems with Selecting Initial Centroids**

- Probability of starting with exactly one initial centroid per ‘real’ cluster is very low
  - K selected for algorithm might be different from inherent K of the data
  - Might randomly select multiple initial objects from same cluster

- Sometimes initial centroids will readjust themselves in the ‘right’ way, and sometimes they don’t

**10 Clusters Example**

- Starting with two initial centroids in one cluster of each pair of clusters
10 Clusters Example

Starting with two initial centroids in one cluster of each pair of clusters

Starting with some pairs of clusters having three initial centroids, while other have only one.

10 Clusters Example

Iteration 1

Iteration 2

Iteration 3

Iteration 4

Starting with some pairs of clusters having three initial centroids, while other have only one.

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 the initial centroids
  - Select those that are most widely separated
- Postprocessing
  - Eliminate small clusters that may represent outliers
  - Split clusters with high SSE
  - Merge clusters that are ‘close’ and have low SSE

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

Limitations of K-means: Differing Sizes

Original Points

K-means (3 Clusters)
Limitations of K-means: Differing Density

Limitations of K-means: Non-globular Shapes

Overcoming K-means Limitations

Overcoming K-means Limitations

Overcoming K-means Limitations

K-Means and Outliers

- K-means algorithm is sensitive to outliers
  - Centroid is average of cluster members
  - Outlier can dominate average computation
- Solution: K-medoids
  - Medoid = most centrally located real object in a cluster
  - Algorithm similar to K-means, but finding medoid is much more expensive
    - Try all objects in cluster to find the one that minimizes SSE, or just try a few randomly to reduce cost
Cluster Analysis Overview

- Introduction
- Foundations: Measuring Distance (Similarity)
- Partitioning Methods: K-Means
- Hierarchical Methods
- Density-Based Methods
- Clustering High-Dimensional Data
- Cluster Evaluation

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Visualized as a dendrogram
  - 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 number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level
- 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 given objects 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 single object (or there are k clusters)

Agglomerative Clustering Algorithm

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

Starting Situation

- Clusters of individual objects, proximity matrix

Proximity Matrix

1 2 3 4 5 6 7 8 9 10
1 0 2 3 4 5 6 7 8 9
2 0 0 1 2 3 4 5 6 7
3 0 0 0 0 1 2 3 4 5
4 0 0 0 0 0 0 1 2 3
5 0 0 0 0 0 0 0 1 2
6 0 0 0 0 0 0 0 0 1
7 0 0 0 0 0 0 0 0 0
8 0 0 0 0 0 0 0 0 0
9 0 0 0 0 0 0 0 0 0
10 0 0 0 0 0 0 0 0 0
Intermediate Situation
• Some clusters are merged

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>C2</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Proximity Matrix

Intermediate Situation
• Merge closest clusters (C2 and C5) and update proximity matrix

After Merging
• How do we update the proximity matrix?

<table>
<thead>
<tr>
<th>C2 U C5</th>
<th>C2</th>
<th>C1</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Proximity Matrix

Defining Cluster Distance
• Min: clusters near each other
• Max: low diameter
• Avg: more robust against outliers
• Distance between centroids

Strength of MIN
• Can handle non-elliptical shapes

Limitations of MIN
• Sensitive to noise and outliers
Strength of MAX

- Less susceptible to noise and outliers

Limitations of MAX

- Tends to break large clusters
- Biased towards globular clusters

Hierarchical Clustering: Average

- Compromise between Single and Complete Link
- Strengths
  - Less susceptible to noise and outliers
- Limitations
  - Biased towards globular clusters

Cluster Similarity: Ward’s Method

- Distance of two clusters is based on the increase in squared error when two clusters are merged
  - Similar to group average if distance between objects is distance squared
- Less susceptible to noise and outliers
- Biased towards globular clusters
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

Time and Space Requirements

- $O(n^2)$ space for proximity matrix
  - $n =$ number of objects
- $O(n^3)$ time in many cases
  - There are $n$ steps and at each step the proximity matrix must be updated and searched
  - Complexity can be reduced to $O(n^2 \log(n))$ time for some approaches
Hierarchical Clustering: Problems and Limitations

- Once a decision is made to combine two clusters, it cannot be undone.
- No objective function is directly minimized.
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers.
  - Difficulty handling different sized clusters and convex shapes.
  - Breaking large clusters.

Cluster Analysis Overview

- Introduction
- Foundations: Measuring Distance (Similarity)
- Partitioning Methods: K-Means
- Hierarchical Methods
- Density-Based Methods
- Clustering High-Dimensional Data
- Cluster Evaluation

Density-Based Clustering Methods

- Clustering based on density of data objects in a neighborhood.
  - Local clustering criterion.
- Major features:
  - Discover clusters of arbitrary shape.
  - Handle noise.
  - Need density parameters as termination condition.

DBSCAN: Basic Concepts

- Two parameters:
  - Eps: Maximum radius of the neighborhood.
    - \( N_{\text{eps}}(q): \{ p \in D | \text{dist}(q,p) \leq \text{Eps} \} \)
  - MinPts: Minimum number of points in an Eps-neighborhood of that point.
- A point p is directly density-reachable from a point q w.r.t. Eps and MinPts if:
  - p belongs to \( N_{\text{eps}}(q) \).
  - Core point condition: \( |N_{\text{eps}}(q)| \geq \text{MinPts} \).

DBSCAN: Classes of Points

- A point is a core point if it has more than a specified number of points (MinPts) within Eps.
  - At the interior of a cluster.
- A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point.
  - At the outer surface of a cluster.
- A noise point is any point that is not a core point or a border point.
  - Not part of any cluster.

Density-Reachable, Density-Connected

- A point p is density-reachable from a point q w.r.t. Eps, MinPts if there is a chain of points
  \( q = p_0, p_1, \ldots, p_n = p \)
  such that \( p_{i+1} \) is directly density-reachable from \( p_i \).
- A point p is density-connected to a point q w.r.t. Eps, MinPts if there is a point o such that both p and q are density-reachable from o w.r.t. Eps and MinPts.
- Cluster is set of density-connected points.
DBSCAN Algorithm

- Repeat until all points have been processed
  - Select a point p
  - If p is core point then
    - Retrieve and remove all points density-reachable from p w.r.t. Eps and MinPts; output them as a cluster
- "Discards" all noise points (how?)
- Discovers clusters of arbitrary shape
- Fairly robust against noise
- Runtime: $O(n^2)$, space: $O(n)$
  - $O(n \times \text{timeToFindPointsInNeighborhood})$
  - Can be $O(n \log(n))$ with spatial index

DBSCAN: Core, Border, and Noise Points

Original Points

Point types: core, border and noise

Eps = 10, MinPts = 4

When DBSCAN Works Well

Original Points

Clusters

When DBSCAN Does NOT Work Well

Varying densities
High-dimensional data

(MinPts=4, large Eps)

(MinPts=4, small Eps)

DBSCAN: Determining Eps and MinPts

- Idea: for points in a cluster, their k-th nearest neighbors are at roughly the same distance
  - Noise points have the k-th nearest neighbor at farther distance
- Plot the sorted distance of every point to its k-th nearest neighbor
  - Choose Eps where sharp change occurs
  - MinPts = k
- k too large: small clusters labeled as noise
- k too small: small groups of outliers labeled as cluster

DBSCAN: Core, Border, and Noise Points
Cluster Analysis Overview

- Introduction
- Foundations: Measuring Distance (Similarity)
- Partitioning Methods: K-Means
- Hierarchical Methods
- Density-Based Methods
  - Clustering High-Dimensional Data
  - Cluster Evaluation

Clustering High-Dimensional Data

- Many applications: text documents, DNA micro-array data
- Major challenges:
  - Irrelevant dimensions may mask clusters
  - Curse of dimensionality for distance computation
  - Clusters may exist only in some subspaces
- Methods
  - Feature transformation, e.g., PCA and SVD
    - Some useful only when features are highly correlated/redundant
    - Feature selection: wrapper or filter approaches
  - Subspace-clustering: find clusters in all subspaces
    - CLIQUE

Curse of Dimensionality

- Graphs on the right adapted from Parsons et al. KDD Explorations ’04
- Data in only one dimension is relatively packed
- Adding a dimension “stretches” the objects across that dimension, moving them further apart
  - High-dimensional data is very sparse
- Distance measure becomes meaningless
  - For many distributions, distances between objects become more similar in high dimensions

Why Subspace Clustering?

- Adapted from Parsons et al. SIGKDD Explorations ’04

CLIQUE (Clustering In QUEst)

- Automatically identifies clusters in sub-spaces
- Exploits monotonicity property
  - If a set of points forms a dense cluster in d dimensions, they also form a cluster in any subset of these dimensions
  - A region is dense if the fraction of data points in the region exceeds the input model parameter \( \varepsilon \)
    - Sound familiar? Apriori algorithm...
- Algorithm is both density-based and grid-based
  - Partitions each dimension into the same number of equal-length intervals
  - Partitions an m-dimensional data space into non-overlapping rectangular units
  - Cluster = maximal set of connected dense units within a subspace
CLIQUE Algorithm

• Find all dense regions in 1-dim space for each attribute. This is the set of dense 1-dim cells. Let k=1.

• Repeat until there are no dense k-dim cells
  – k = k+1
  – Generate all candidate k-dim cells from dense (k-1)-dim cells
  – Eliminate cells with fewer than \( \zeta \) points

• Find clusters by taking union of all adjacent, high-density cells of same dimensionality

• Summarize each cluster using a small set of inequalities that describe the attribute ranges of the cells in the cluster

Strengths and Weaknesses of CLIQUE

• Strengths
  – Automatically finds subspaces of the highest dimensionality that contain high-density clusters
  – Insensitive to the order of objects in input and does not presume some canonical data distribution
  – Scales linearly with input size and has good scalability with number of dimensions

• Weaknesses
  – Need to tune grid size and density threshold
  – Each point can be a member of many clusters
  – Can still have high mining cost (inherent problem for subspace clustering)
  – Same density threshold for low and high dimensionality

Cluster Analysis Overview

• Introduction
• Foundations: Measuring Distance (Similarity)
• Partitioning Methods: K-Means
• Hierarchical Methods
• Density-Based Methods
• Clustering High-Dimensional Data
• Cluster Evaluation

Cluster Validity on Test Data

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Entertainment</th>
<th>Financial</th>
<th>Foreign</th>
<th>Metro</th>
<th>National</th>
<th>Sports</th>
<th>Entropy</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>40</td>
<td>206</td>
<td>94</td>
<td>27</td>
<td>1.2770</td>
<td>0.7474</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
<td>290</td>
<td>20</td>
<td>33</td>
<td>21</td>
<td>1.1472</td>
<td>0.7956</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>671</td>
<td>4</td>
<td>0.1813</td>
<td>0.9796</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>162</td>
<td>2</td>
<td>3</td>
<td>119</td>
<td>73</td>
<td>1.7487</td>
<td>0.4090</td>
</tr>
<tr>
<td>5</td>
<td>331</td>
<td>22</td>
<td>3</td>
<td>70</td>
<td>13</td>
<td>23</td>
<td>1.5976</td>
<td>0.7334</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>358</td>
<td>12</td>
<td>212</td>
<td>44</td>
<td>13</td>
<td>1.0023</td>
<td>0.5325</td>
</tr>
<tr>
<td>Total</td>
<td>1034</td>
<td>355</td>
<td>341</td>
<td>943</td>
<td>273</td>
<td>708</td>
<td>1.1430</td>
<td>0.7260</td>
</tr>
</tbody>
</table>

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

\[
p_{cj} = \frac{n_{cj}}{n_{c}},
\]

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

\[
\text{Entropy} = -\sum_{c=1}^{C} p_{cj} \log_{2} p_{cj},
\]

where \( C \) 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 the cluster, i.e.,

\[
\text{Total Entropy} = \sum_{j=1}^{K} n_{j} \cdot \text{Entropy}.
\]

**Purity** Using the terminology derived for entropy, the purity of cluster \( j \) is given by

\[
purity_{j} = \frac{m_{j}}{n_{j}},
\]

where \( m_{j} \) is the size of cluster \( j \) and \( n_{j} \) is the total number of data points.

\[
purity_{j} = \frac{m_{j}}{n_{j}}
\]

Cluster Validity

• Clustering: usually no ground truth available
• Problem: “clusters are in the eye of the beholder…”

• Then why do we want to evaluate them?
  – To avoid finding patterns in noise
  – To compare clustering algorithms
  – To compare two sets of clusters
  – To compare two clusters
Clusters found in Random Data

Measuring Cluster Validity Via Correlation

- Two matrices
  - Similarity Matrix
  - "Incidence" Matrix
    - One row and one column for each object
    - Entry is 1 if the associated pair of objects belongs to the same cluster, otherwise 0
- Compute correlation between the two matrices
  - Since the matrices are symmetric, only the correlation between n(n-1) / 2 entries needs to be calculated.
- High correlation: objects close to each other tend to be in same cluster
- Not a good measure when clusters can be non-globular and intertwined

Similarity Matrix for Cluster Validation

- Order the similarity matrix with respect to cluster labels and inspect visually
  - Block-diagonal matrix for well-separated clusters

Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp
Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp

![Complete Link](image)

DBSCAN

Sum of Squared Error

- For fixed number of clusters, lower SSE indicates better clustering
  - Not necessarily true for non-globular, intertwined clusters
- Can also be used to estimate the number of clusters
  - Run K-means for different K, compare SSE

![SSE of clusters found using K-means](image)

Comparison to Random Data or Clustering

- Need a framework to interpret any measure
  - E.g., if measure = 10, is that good or bad?
- Statistical framework for cluster validity
  - Compare cluster quality measure on random data or random clustering to those on real data
    - If value for random setting is unlikely, then cluster results are valid (cluster = non-random structure)
  - For comparing the results of two different sets of cluster analyses, a framework is less necessary
    - But: need to know whether the difference between two index values is significant

Statistical Framework for SSE

- Example: found 3 clusters, got SSE = 0.005 for given data set
- Compare to SSE of 3 clusters in random data
  - Histogram: SSE of 3 clusters in 500 sets of random data points (100 points from range 0.2…0.8 for x and y)
  - Estimate mean, stdv for SSE on random data
  - Check how many stdv away from mean the real-data SSE is

![SSE of clusters found using K-means](image)
Statistical Framework for Correlation

- Compare correlation of incidence and proximity matrices for well-separated data versus random data

Cluster Cohesion and Separation

- **Cohesion**: how closely related are objects in a cluster
  - Can be measured by SSE ($m_i$ = centroid of cluster $i$):
    $$\text{SSE} = \sum \sum (x - m)^2 = \sum \frac{1}{2} \sum (x_i - y_i)^2$$

- **Separation**: how well-separated are clusters
  - Can be measured by between-cluster sum of squares ($m = \text{overall mean}$):
    $$\text{BSS} = \sum C_i (m - m^2)$$

Cohesion and Separation Example

- **Note**: BSS + SSE = constant
  - Minimize SSE => get max. BSS

Silhouette Coefficient

- Combines ideas of both cohesion and separation
- For an individual object $i$
  - Calculate $a_i = \text{average distance of } i \text{ to the objects in its cluster}$
  - Calculate $b_i = \text{average distance of } i \text{ to objects in another cluster } C$, choosing the $C$ that minimizes $b_i$
  - Silhouette coefficient of $i = (b_i - a_i) / \max\{a_i, b_i\}$
    - Range: [-1,1], but typically between 0 and 1
    - The closer to 1, the better
- Can calculate the Average Silhouette width over all objects

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
Summary

- Cluster analysis groups objects based on their similarity (or distance) and has wide applications.
- Measure of similarity (or distance) can be computed for all types of data.
- Many different types of clustering algorithms
  - Discover different types of clusters.
- Many measures of clustering quality, but absence of ground truth always a challenge.