Clustering Algorithms and Validity measures

— Tutorial —

Maria Halkidi

Dept. of Informatics
Athens Univ. of Economics & Business,
email: mhalk@aueb.gr
http://www.db-net.aueb.gr

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
CONTENTS

_clusters_ – Unsupervised Learning

✓ What is Cluster Analysis?
✓ Clustering Applications
✓ Clustering Algorithms

_clusters_ Quality Assessment

✓ Validity Criteria
✓ Cluster Validity Approaches

_clusters_ Trends in Clustering

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Clustering
(Unsupervised Learning)
Cluster analysis aims at:
- grouping a set of data objects into clusters
- identifying interesting distributions and patterns in underlying data

Cluster: A collection of data objects that are
- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters

Clustering is perceived as an unsupervised learning procedure:
- no predefined classes and no examples
  that would show what kind of desirable relations should be valid among the data

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Steps of a clustering process

- **Feature selection.** Select properly the features on which clustering is to be performed.

- **Clustering algorithm.**
  - Proximity measure.
  - Clustering criterion

- **Validation of the results.** The correctness of clustering algorithm results is verified using appropriate criteria and techniques.

- **Interpretation of the results.** The experts in the application area integrate the clustering results with other experimental evidence and analysis in order to draw the right conclusion.
Clustering Directions

- **Data reduction.** Instead of processing the data set as an entity, we adopt the representatives of the defined clusters in our process. Thus, data compression is achieved.

- **Hypothesis generation.** Cluster analysis is used here in order to infer some hypotheses concerning the data.

**Clustering applied to a retail database.** → two significant groups

**Defined hypothesis:** “young people go shopping in the evening”, “old people go shopping in the morning”.

- **Prediction based on groups.** Unknown patterns can be classified into specified clusters based on their similarity to the clusters’ features. Useful knowledge related to our data can be extracted.
Clustering Application Domains

- **Business**.
  - Discover significant groups of customers and characterize them based on their purchasing patterns

- **Biology**.
  - Categorize genes with similar functionality,
  - Gain insight into gene structures

- **Spatial Data Analysis**
  - detect spatial clusters and automate the process of large spatial databases

- **WWW**
  - Document classification
  - Cluster Web-log data to discover groups of users with similar access patterns

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
A Categorization of Clustering Methods

- **Technique used** in order to define clusters.
  - Partitional Methods
  - Hierarchical Methods
  - Density-Based Methods
  - Grid-Based Methods

- **The type of variables**
  - Statistical – Numerical Data
  - Conceptual- Categorical Data

- **Theory used in order to extract clusters**
  - Fuzzy Clustering
  - Crisp Clustering
  - Kohonen net clustering
Partitional Algorithms: Basic Concept

- **Partitional method:** Decompose the data set into a set of k disjoint clusters.

- **Problem Definition:**
  - Given an integer k,
  - Find a partition of k clusters that optimizes the chosen partitioning criterion
The $K$-Means Clustering Method

\[ E_K = \sum_k \left\| x_k - m_{c(x_k)} \right\|^2 \]
Variations of the *K-Means* Method

- A few variants of the *k-means* which differ in
  - Selection of the initial $k$ centers
  - Dissimilarity calculations
  - Strategies to calculate cluster centers

- Handling categorical data: *k-modes*
  - Replacing cluster centers with *modes*
  - Using new dissimilarity measures to deal with categorical objects
  - A mixture of categorical and numerical data: *k-prototype* method
PAM (Partitioning Around Medoids)

- **Step 1**: Select k representative objects (medoids) arbitrarily
- **Step 2**: Compute **Total Swapping Cost** for all pairs of objects $O_i, O_h$ where $O_i$ is currently selected, and $O_h$ is not
  \[ TC_{ih} = \sum_j C_{jih} \]
- **Step 3**: Select the pair $O_i, O_h$ which minimizes $TC_{ih}$.
- **Step 4**: If $\min TC_{ih} < 0$, replace $O_i$ with $O_h$ and go to step 2
  else
  for each non-selected object find the most similar representative object. Halt
CLARA
(Clustering LARge Applications)

- It draws multiple samples of the data set, applies PAM to each sample, and gives the best clustering as the output
- **Advantage:** deals with larger data sets than PAM
- **Problems:**
  - Efficiency depends on the sample size
  - A good clustering based on samples will not necessarily represent a good clustering of the whole data set if the sample is biased
CLARANS ("Randomized" CLARA)

- The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of $k$ medoids.
- It selects a node and compares it to a user-defined number of their neighbors searching for local minimum.
- If the local optimum is found, CLARANS starts with new randomly selected node in search for a new local optimum.

**Input**: number of local minimum, number of neighbors

- CLARANS draws sample of neighbors dynamically.
- It is more efficient and scalable than both PAM and CLARA.
Hierarchical Clustering

*Hierarchical clustering* proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The result of the algorithm is a tree of clusters, called *dendrogram*.
Hierarchical Clustering Algorithms

- **BIRCH (1996):** uses CF-tree and incrementally adjusts the quality of sub-clusters.

- **CURE (1998):** is robust to outliers and identifies clusters of non-spherical shapes.

- **ROCK (1999):** is a robust clustering algorithm for Boolean and categorical data. It introduces two new concepts, that is a point's neighbours and links.
BIRCH (1996)
(Balanced Iterative Reducing and Clustering using Hierarchies)

- Incrementally construct a **CF (Clustering Feature) tree**, a hierarchical data structure for multiphase clustering

  ✓ **Phase 1**: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)

  ✓ **Phase 2**: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree

- **Scales linearly**: finds a good clustering with a single scan and improves the quality with a few additional scans

- **Problem**: handles only numeric data, and sensitive to the order of the data record.

---

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Clustering Feature Vector

Clustering Feature: $CF = (N, LS, SS)$

$N$: Number of data points

$LX$: $\sum_{i=1}^{N} X_i$

$SS$: $\sum_{i=1}^{N} X_i^2$

$CF = (5, (16,30),(54,190))$

\[(3,4)\]  
\[(2,6)\]  
\[(4,5)\]  
\[(4,7)\]  
\[(3,8)\]
CF Tree

Root

<table>
<thead>
<tr>
<th>CF1</th>
<th>CF2</th>
<th>CF3</th>
<th>...</th>
<th>CF6</th>
</tr>
</thead>
<tbody>
<tr>
<td>child1</td>
<td>child2</td>
<td>child3</td>
<td>...</td>
<td>child6</td>
</tr>
</tbody>
</table>

Non-leaf node

<table>
<thead>
<tr>
<th>CF1</th>
<th>CF2</th>
<th>CF3</th>
<th>...</th>
<th>CF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>child1</td>
<td>child2</td>
<td>child3</td>
<td>...</td>
<td>child5</td>
</tr>
</tbody>
</table>

Leaf node

<table>
<thead>
<tr>
<th>prev</th>
<th>CF1</th>
<th>CF2</th>
<th>...</th>
<th>CF6</th>
<th>next</th>
</tr>
</thead>
</table>

| prev | CF1 | CF2 | ... | CF4 | next |

B = 7 (non-leaf entries),
T = radius or diameter of leaf entries,
L = 6 (leaf entries)

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
CURE
(Clustering Using REpresentatives)

• **Drawbacks of partitional and hierarchical clustering methods**
  
  ✓ Consider only one point as representative of a cluster
  
  ✓ Good only for convex shaped, similar size and density, and if *number of clusters* can be reasonably estimated

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
CURE

- Identifies clusters having non-spherical shape.
  - Uses multiple representative points to evaluate the distance between clusters, adjusts well to arbitrary shaped clusters

- Handles outliers.
  - Shrink the multiple representative points towards the center of cluster by a fraction, $\alpha$.

- **CURE** employs a combination of
  - Random sampling
  - Partitioning

- **Problems**: Priori-known number of clusters, $k$.

  *Time complexity*: $O(n^2 \log n)$

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Clustering Categorical Data: ROCK

- **ROCK:**
  - Handles Boolean and Categorical data

- **Basic ideas:**
  - **Similarity function:**
    \[
    Sim(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}
    \]
  - **Neighbors:**
    \[
    Sim(T_1, T_2) \geq \theta \quad \Rightarrow \quad T_1 \text{ and } T_2 \text{ neighbors}
    \]
  - **Links:** The number of common neighbors for the two points.

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Density-Based Clustering Algorithms

- **Clustering** based on density (local cluster criterion), such as density-connected points

- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - Need density parameters as termination condition

- **Representative algorithms:**
  - **DBSCAN**: Ester, et al.
  - **DENCLUE**: Hinneburg & D. Keim
Density-Based Clustering: Background

- Two parameters:
  - \( Eps \): Maximum radius of the neighborhood
  - \( MinPts \): Minimum number of points in an Eps-neighborhood of that point

- \( N_{Eps}(p) \): \( \{ q \in D \mid \text{dist}(p,q) \leq Eps \} \)

- Directly density-reachable:
  - \( p \) belongs to \( N_{Eps}(q) \)
  - Core point condition:
    \[ |N_{Eps}(q)| \geq MinPts \]

\[ MinPts = 5 \]
\[ Eps = 1 \text{ cm} \]
DBSCAN

- A cluster is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise

![Diagram showing clusters, core, border, and outlier points with Eps = 1cm and MinPts = 5]
DBSCAN: The Algorithm

- Arbitrary select a point $p$
- Retrieve all points density-reachable from $p$ wrt $\text{Eps}$ and $\text{MinPts}$.
- If $p$ is a core point, a cluster is formed.
- If $p$ is a border point, no points are density-reachable from $p$ and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.
DENCLUE

- Uses grid cells. It keeps information about grid cells that do actually contain data points

- The basic idea is to model the overall density of data space as the sum of influence functions of the data points.
  - Influence function: describes the impact of a data point within its neighbourhood (square wave function, Gaussian,…)
  - Density function: Sum of the influences of all data points

Clusters can be identified by determining density attractors.

Density attractors are local maximum of the overall density function.
Grid - Based Clustering Method

- Quantize the space into a finite number of cells and then do all operations on the quantized space

- Several interesting methods
  - **STING** (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - **WaveCluster** by Sheikholeslami, Chatterjee, and Zhang (VLDB’98)
STING
(Statistical Information Grid-based method)

- It divides the spatial area into rectangular cells
- It computes the statistical parameters (such as mean, variance, minimum, maximum and type of distribution) of each numerical feature of the objects within cells.
- It generates a hierarchical structure of the grid cells so as to represent the clustering information at different levels.
- The hierarchical structure of STING enables the usage of clustering information to search for queries or the efficient assignment of a new object to the clusters.
WaveCluster

- Clustering from a signal processing perspective using wavelets.

- Input parameters:
  - number of grid cells for each dimension
  - the wavelet, and
  - the number of applications of wavelet transform.

**Problem:** Does not work in high dimensional space
WaveCluster : The Algorithm

- Spatial data is represented in a n-dimensional feature space.

**Basic Idea:** Apply wavelet-transformation to the feature space

- A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-bands.

  - **High frequency parts of signal** $\Rightarrow$ **Boundaries of clusters**
  - **Low frequency** $\Rightarrow$ **Clusters**

- Find the connected components as clusters in the average sub-bands of transformed feature space, at different levels

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Fuzzy Clustering

- **Crisp clustering**, meaning that a data point either belongs to a class or not.
- **Fuzzy Clustering** a data point may belong to more than one clusters with different degrees of belief.

Representative fuzzy clustering algorithm: **Fuzzy C-Means (FCM).**

**FCM objective function:**

\[
J_m(U,V) = \sum_{i=1}^c \sum_{k=1}^n U_{ik}^m d^2(x_k, v_i)
\]

- \( m \to 1 \Rightarrow \text{clusters} \to \text{crisp} \)
- \( m \to \infty \Rightarrow \text{clusters} \to \text{fuzzy}, U_{ik} \to 1/c \)
<table>
<thead>
<tr>
<th>Category</th>
<th>Partitional</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Type of data</td>
<td>Complexity</td>
<td>Geometry</td>
<td>Outliers, noise</td>
<td>Input parameters</td>
<td>Results</td>
</tr>
<tr>
<td>K-Mean</td>
<td>Numerical</td>
<td>O(n)</td>
<td>non-convex shapes</td>
<td>No</td>
<td>Number of clusters</td>
<td>Center of clusters</td>
</tr>
<tr>
<td>K-mode</td>
<td>Categorical</td>
<td>O(n)</td>
<td>non-convex shapes</td>
<td>No</td>
<td>Number of clusters</td>
<td>Modes of clusters</td>
</tr>
<tr>
<td>PAM</td>
<td>Numerical</td>
<td>O(k(n-k)^2)</td>
<td>non-convex shapes</td>
<td>No</td>
<td>Number of clusters</td>
<td>Medoids of clusters</td>
</tr>
<tr>
<td>CLARA</td>
<td>Numerical</td>
<td>O(k(40+k)^2 + k(n-k))</td>
<td>non-convex shapes</td>
<td>No</td>
<td>Number of clusters</td>
<td>Medoids of clusters</td>
</tr>
<tr>
<td>CLARANS</td>
<td>Numerical</td>
<td>O(kn^2)</td>
<td>non-convex shapes</td>
<td>No</td>
<td>Number of clusters, maximum number of neighbors examined</td>
<td>Medoids of clusters</td>
</tr>
<tr>
<td>FCM</td>
<td>Numerical</td>
<td>O(n)</td>
<td>Non-convex shapes</td>
<td>No</td>
<td>Number of clusters</td>
<td>Center of cluster, beliefs</td>
</tr>
</tbody>
</table>

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
<table>
<thead>
<tr>
<th>Name</th>
<th>Type of data</th>
<th>Complexity*</th>
<th>Geometry</th>
<th>Outliers</th>
<th>Input parameter</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIRCH</td>
<td>Numerical</td>
<td>O(n)</td>
<td>non-convex shapes</td>
<td>Yes</td>
<td>Radius of clusters, branching factor</td>
<td>CF = (N, LS, SS)</td>
</tr>
<tr>
<td>CURE</td>
<td>Numerical</td>
<td>O(n²logn)</td>
<td>Arbitrary shapes</td>
<td>Yes</td>
<td>Number of clusters, number of clusters representatives</td>
<td>Assignment of data values to clusters</td>
</tr>
<tr>
<td>ROCK</td>
<td>Categorical</td>
<td>O(n²+nm_m_a + n²logn), O(n², nm_m_a)</td>
<td>Arbitrary shapes</td>
<td>Yes</td>
<td>Number of clusters</td>
<td>Assignment of data values to clusters</td>
</tr>
</tbody>
</table>

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Type of data</th>
<th>Complexity</th>
<th>Geometry</th>
<th>Outliers</th>
<th>Input parameters</th>
<th>Results</th>
<th>Clustering criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DENCLU</td>
<td>Numerical</td>
<td>O(nlogn)</td>
<td>Arbitrary shapes</td>
<td>Yes</td>
<td>Cluster radius, minimum number of objects</td>
<td>Assignment of data values to clusters</td>
<td>Merge points that are density reachable into one cluster.</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>Numerical</td>
<td>O(logn)</td>
<td>Arbitrary shapes</td>
<td>Yes</td>
<td>Cluster radius $\sigma$, Minimum number of objects $\xi$</td>
<td>Assignment of data values to clusters</td>
<td>$x^<em>$ density attractor for a point $x$ if $F_{Gauss} &gt; \xi$ then $x$ attached to the cluster belonging to $x^</em>$.</td>
</tr>
<tr>
<td>Category</td>
<td>Grid-Based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Name</strong></td>
<td><strong>Type of data</strong></td>
<td><strong>Complexity</strong></td>
<td><strong>Geometry</strong></td>
<td><strong>Outliers</strong></td>
<td><strong>Input parameters</strong></td>
<td><strong>Output</strong></td>
<td><strong>Clustering criterion</strong></td>
<td></td>
</tr>
<tr>
<td>Wave-Cluster</td>
<td>Spatial data</td>
<td>O(n)</td>
<td>Arbitrary shapes</td>
<td>Yes</td>
<td>Wavelets, the number of grid cells for each dimension, the number of applications of wavelet transform.</td>
<td>Clustersed objects</td>
<td>Decompose feature space applying wavelet transformation <em>Average sub-band</em> → clusters <em>Detail sub-bands</em> → clusters boundaries</td>
<td></td>
</tr>
<tr>
<td>STING</td>
<td>Spatial data</td>
<td>O(K) <em>K is the number of grid cells at the lowest level</em></td>
<td>Arbitrary shapes</td>
<td>Yes</td>
<td>Number of objects in a cell</td>
<td>Clustersed objects</td>
<td>Divide the spatial area into rectangle cells and employ a hierarchical structure. Each cell at a high level is partitioned into a number of smaller cells in the next lower level</td>
<td></td>
</tr>
</tbody>
</table>

M. Halkidi, Y. Batistakis, M. Vaziriannis, SSDBM, July 2001
Clustering Quality Assessment

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Cluster Validity -
Problem Specification

A problem we face in clustering is to decide the optimal number of clusters that fits a data set.

The various clustering algorithms behave in a different way depending on:

i) the features of the data set (geometry and density distribution of clusters)

ii) the input parameters values
Clustering Quality Assessment

What is Good Clustering?

✓ “How many clusters are there in the data set?”
✓ “Does the defined clustering scheme fits our data set?”
✓ “Is there a better clustering possible?”
Fundamental concepts of cluster validity

The procedure of evaluating the results of a clustering algorithm is known under the term *cluster validity*.

Three approaches to investigate cluster validity:

- **External criteria.** We evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set.

- **Internal criteria.** We may evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves (e.g. proximity matrix).

- **Relative criteria.** The basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different parameter values.
Clustering Quality Criteria

- **high intra-cluster similarity**
  - Variance

- **low inter-cluster similarity**
  - Single Linkage
  - Complete Linkage
  - Comparison of centroids
Cluster validity approaches

External and Internal approaches

✓ based on statistical tests → high computational cost
✓ indices related to these approaches aim at measuring the degree to which a data set confirms an a-priori specified scheme.

Relative approaches

✓ finding the best clustering scheme that a clustering algorithm can define under certain assumptions and parameters.
External Criteria

The basic idea is to test whether the points of the data set are randomly structured or not.

This analysis is based on the *Null Hypothesis, Ho*, expressed as a statement of random structure of a dataset.

Based on the external criteria we can work in two different ways:
- Comparison of clustering structure $C$ with partitioning $P$
- Comparison of proximity matrix $P$ with partitioning $P$
Consider $C = \{C_1\ldots C_m\}$ is a clustering structure of a data set $X$ and $P = \{P_1\ldots P_s\}$ is a defined partition of the data.

We refer to a pair of points $(x_v, x_u)$ from the data set using the following terms:

- **SS**: if both points belong to the same cluster of the clustering structure $C$ and to the same group of partition $P$.
- **SD**: if points belong to the same cluster of $C$ and to different groups of $P$.
- **DS**: if points belong to different clusters of $C$ and to the same group of $P$.
- **DD**: if both points belong to different clusters of $C$ and to different groups of $P$. 
Comparison of $C$ with partition $P$

We can define the following indices to measure the degree of similarity between $C$ and $P$:

- **Rand Statistic:** $R = (a + d) / M$,
- **Jaccard Coefficient:** $J = a / (a + b + c)$,

where $a$, $b$, $c$, and $d$ are the number of SS, SD, DS, and DD pairs respectively.

- $a + b + c + d = M$ which is the maximum number of all pairs in the data set,
- $M = N(N-1)/2$ where $N$ is the total number of pairs of points in the data set.
External Validity Indices

- **Folkes and Mallows index:**
  \[ FM = \frac{a}{\sqrt{m_1 m_2}} = \frac{a}{\sqrt{\frac{a + b}{a + c} \cdot \frac{a}{a + c}}} \]
  where \( m_1 = \frac{a}{a + b} \), \( m_2 = \frac{a}{a + c} \).

- **Huberts \( \Gamma \) statistic:**
  \[ \Gamma = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i, j) Y(i, j) \]
  high values of indices indicate great similarity between \( C \) and \( P \).

- **Normalized \( \Gamma \) statistic**
  \[ \bar{\Gamma} = \left[ \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i, j) - \mu_X)(Y(i, j) - \mu_Y) \right] / \sigma_X \sigma_Y \]
  where \( \mu_X, \mu_Y, \sigma_X, \sigma_Y \) are the respective means and variances of \( X, Y \) matrices.
Evaluation Procedure based on External Criteria

Let a data set $X$ and $C = \{C_1, \ldots, C_m\}$ is a clustering structure of $X$ as defined by a clustering algorithm. 

$P = \{P_1, \ldots, P_s\}$ is a defined partition of the data, where $m \neq s$.

For $i = 1$ to $r$

- **Generate** a data set $X_i$ with $N$ vectors (points) in the area of $X$.
- **Assign** each vector $y_{j,i}$ of $X_i$ to the group that $x_j \in X$ belongs, according to the partition $P$.
- **Run** the same clustering algorithm used to produce structure $C$, for each $X_i$, and let $C_i$ the resulting clustering structure.
- **Compute** $q(C_i)$ value of the defined index $q$ for $P$ and $C_i$.

End For

Create scatter-plot of the $r$ validity index values, $q(C_i)$ (that computed into the for loop).

✓ Compare validity index value, let $q$, to the $q(C_i)$ values, let $q_i$.

✓ The indices $R, J, FM, \Gamma$ defined previously are used as the $q$ index mentioned in the above procedure.
Comparison of P (proximity matrix) with partition P

Partition $P$ can be considered as a mapping
$$g : X \rightarrow \{1 \ldots n_c\}.$$ 

Assuming matrix
$$Y : Y(i, j) = \begin{cases} 1, & \text{if } g(x_i) \neq g(x_j) \text{ and} \\ 0, & \text{otherwise} \end{cases}, \quad i, j = 1 \ldots N,$$

- We compute $\Gamma$ (or normalized $\Gamma$) statistic using the proximity matrix $P$ and the matrix $Y$.

  **Index value → an indication of the two matrices’ similarity.**

- To proceed with the evaluation procedure we use the Monde Carlo techniques as mentioned above.

  “Generate” → step of the procedure we generate the corresponding mappings $g_i$ for every generated $X_i$ data set.

  “Compute” → step we compute the matrix $Y_i$, for each $X_i$ in order to find the $\Gamma_i$ corresponding statistic index.
Internal Criteria

We evaluate the clustering result of an algorithm using only quantities and features inherent to the dataset.

Two cases in which we apply internal criteria of cluster validity depending on the clustering structure:

a) hierarchy of clustering schemes, and

b) single clustering scheme.
Validating hierarchy of clustering schemes

Cophenetic matrix, $P_c$ → represent the hierarchy diagram that produced by a hierarchical algorithm

✓ **Cophenetic Correlation Coefficient**: A statistical index to measure the degree of similarity between $P_c$ and $P$ (proximity matrix)

$$CPCC = \frac{\left(\frac{1}{M}\right)\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} c_{ij} - \mu_P \mu_C}{\sqrt{\left(\frac{1}{M}\right)\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 - \mu_P^2} \left[\left(\frac{1}{M}\right)\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} c_{ij}^2 - \mu_C^2\right]}, \quad -1 \leq CPCC \leq 1$$

where $M=N\cdot(N-1)/2$ and $N$ is the number of points in a dataset

$$\mu_P = \left(\frac{1}{M}\right)\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} P(i, j), \quad \mu_C = \left(\frac{1}{M}\right)\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} P_c(i, j)$$

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Validating a single clustering scheme

The goal here is to find the degree of agreement between
- a given clustering scheme $C$, consisting of $k$ clusters, and
- the proximity matrix $P$.

The defined indices for this approach are
- Hubert’s $\Gamma$ statistic
  - or
- normalized $\Gamma$ statistic.

To compute the indices we use:
- An matrix defined as
  $$Y(i, j) = \begin{cases} 
  1, & \text{if } x_i \text{ and } x_j \text{ belong to different clusters, } i, j = 1, \ldots, N. \\
  0, & \text{otherwise}
  \end{cases}$$

- Monde Carlo techniques is the way to test the random hypothesis in a given data set.
Relative Criteria

The fundamental idea is to choose the best clustering scheme of a set of defined schemes according to a pre-specified criterion.

The problem can be stated as follows:

“Let $P$ the set of parameters associated with a specific clustering algorithm (e.g. the number of clusters $nc$). Among the clustering schemes $C_i$, $i=1,\ldots,nc$, defined by a specific algorithm, for different values of the parameters in $P$, choose the one that best fits the data set.”
Relative Criteria (continue…)

There are two approaches for defining the best clustering depending on the behaviour of \( q \) with respect to \( nc \).

The validity index

✓ does not exhibit an increasing or decreasing trend as the number of clusters increases

we seek the maximum (minimum) of index in its the plot with respect of \( nc \)

✓ increase (or decrease) as the number of clusters increases

we search for the values of \( nc \) at which a significant local change in value of the index occurs.

Note: the absence of a knee may be an indication that the data set possesses no clustering structure.
Cluster Validity Indices
--Crisp clustering--

The modified Hubert $\Gamma$ statistic

$$\Gamma = \left(1 / M\right) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} P(i, j) \cdot Q(i, j)$$

where

- $M=N(N-1)/2$,
- $P$ is the proximity matrix of the data set and
- $Q$ is an NXN matrix whose $(i, j)$ element is equal to the distance between the representative points $(v_{ci}, v_{cj})$ of the clusters where the objects $x_i$ and $x_j$ belong.

In the plot of normalized $\Gamma$ versus $nc$, the number of clusters at which a significant increase of normalized $\Gamma$ occurs

Indication of the number of clusters that underlie the data

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Cluster Validity Indices
--Crisp clustering--

Dunn index

\[
D_{nc} = \min_{i=1,..., nc} \left\{ \min_{j=i+1,..., nc} \left( \frac{d(c_i, c_j)}{\max_{k=1,..., nc} diam(c_k)} \right) \right\}
\]

- the dissimilarity function between two clusters \(c_i\) and \(c_j\) defined as

\[
d(c_i, c_j) = \min_{x \in c_i, y \in c_j} d(x, y),
\]

- \(diam(c)\) is the diameter of a cluster, which may be considered as a measure of dispersion of the clusters.

\[
diam(C) = \max_{x, y \in C} d(x, y)
\]
Cluster Validity Indices

--Crisp clustering--

Best clustering scheme $\Rightarrow d(c_i, c_j) \uparrow & \text{diam}(c) \downarrow$

The maximum in the plot of $D_{nc}$ versus the number of clusters can be an indication of the number of clusters that fits the data.

The implications of the Dunn index are:

- the considerable amount of time required for its computation,
- the sensitive to the presence of noise in datasets, since these are likely to increase the values of diam(c)
Cluster Validity Indices

--Crisp clustering--

The Davies-Bouldin (DB) index

A similarity measure $R_{ij}$ between the clusters $C_i$ and $C_j$ is defined based on

- a measure of dispersion of a cluster $C_i$ and
- a dissimilarity measure between two clusters $d_{ij}$.

The $R_{ij}$ index is defined to satisfy the following conditions [Davi 79]:

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- if $s_i = 0$ and $s_j = 0$ then $R_{ij} = 0$
- if $s_j > s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik}$
- if $s_j = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$.

A simple choice for $R_{ij}$ that satisfies the above conditions is:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Cluster Validity Indices
--Crisp clustering--

The DB index is defined as

\[ DB_{nc} = \frac{1}{n_c} \sum_{i=1}^{nc} R_i \]

\[ R_i = \max_{i=1,...,nc,i \neq j} R_{ij}, \ i = 1,...,nc \]

☑ DB\(_{nc}\) is the average similarity between each cluster \(c_i\), \(i=1,...,nc\) and its most similar one.

☑ It is desirable for the clusters to have the minimum possible similarity to each other.

☑ The DB\(_{nc}\) index exhibits no trends with respect to the number of clusters.

we seek the minimum value of DB\(_{nc}\) in its plot versus the number of clusters.

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Cluster Validity Indices
--Crisp clustering--

RMSSDT, SPR, RS, CD  \textit{(Hierarchical Clustering Algorithms)}

These four indices can be applied to each step of a \textit{hierarchical} clustering algorithm and they are known as:

- Root-mean-square standard deviation (RMSSTD) of the new cluster
- Semi-partial R-squared (SPR)
- R-squared (RS)
- Distance between two clusters (CD)

They have to be used simultaneously to determine the number of clusters existing in our data set.
Cluster Validity Indices
--Crisp clustering --

➢ **RMSSDT** (*Hierarchical Clustering Algorithms*)

*RMSSTD* is the square root of the attributes used in the clustering process.

✗ It measures the homogeneity of the formed clusters at each step of the hierarchical algorithm.

✗ The RMSSTD of a cluster should be as small as possible.

✗ If the values of RMSSTD are higher at one step than the ones of the previous step, we have an indication that the new clustering scheme is not homogenous.
Cluster Validity Indices
--Crisp clustering--

**SPR (Hierarchical Clustering Algorithms)**

We define the term *Sum of Squares* as

$$SS = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Also we use the following symbolisms:

- $SS_w$ referring to the within group sum of squares,
- $SS_b$ referring to the between groups sum of squares
- $SS_t$ referring to the total sum of squares, of the whole data set.

$$SPR = (SS_w \text{ of the new cluster} - \text{the sum of } SS_w \text{ of clusters joined to obtain the new cluster}) / SS_t \text{ for the whole data set.}$$

This index measures the *loss of homogeneity* after merging the two clusters of a single algorithm step.

- ✓ $SPR = 0$ → the new cluster is obtained by merging two perfectly homogeneous clusters.
- ✓ $SPR > 0$ → the new cluster is obtained by merging two heterogeneous clusters.
Cluster Validity Indices
--Crisp clustering--

- **RS** *(Hierarchical Clustering Algorithms)*

\[ RS = \frac{SS_b}{SS_t} \]

- \( SS_b \) is a measure of difference between groups.
  \[ SS_t = SS_b + SS_w \]

- RS may be considered as a measure of
  - the degree of difference between clusters
  - the degree of homogeneity between groups.

- **RS** = 0 \( \Rightarrow \) no difference exists among groups
- **RS** = 1 \( \Rightarrow \) there is significant difference among groups.

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Cluster Validity Indices
--Crisp clustering--

- **CD** *(Hierarchical Clustering Algorithms)*

  The CD index measures the distance between the two clusters that are merged in a given step.

- **Centroid hierarchical clustering**
  - CD is the distance between the centers of the clusters.

- **Single linkage**
  - CD measures the minimum Euclidean distance between all possible pairs of points

- **Complete linkage**
  - CD is the maximum Euclidean distance between all pairs of data points.
Cluster Validity Indices
--Crisp clustering --

RMSSTD & RS (Non- Hierarchical Clustering Algorithms)

\[
\text{RMSSTD} = \left( \sum_{i=1}^{n} \left( \sum_{k=1}^{n_{ij}} \frac{(x_{kj} - \overline{x}_{kj})^2}{\sum_{i=1}^{n_{ij}} (n_{ij} - 1)} \right) \right)^{1/2}
\]

\[
\text{RS} = \frac{SS_b}{SS_t} = \frac{SS_t - SS_w}{SS_t}
\]

Find the optimal Clustering

✓ Run the algorithm a number of times for different number of clusters each time.
✓ Plot the respective graphs of the validity indices vs number of clusters
✓ Search for the significant “knee” in these graphs.

Optimal clustering for our data set → number of clusters at which the “knee” is observed

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Cluster Validity Indices

--Crisp clustering--

**SD Index**

- **Variance of data set**
  \[ \sigma_x^p = \frac{1}{n} \sum_{k=1}^{n} \left( x_k^p - \bar{x}^p \right)^2 \]
  where \( \bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k, \forall x_k \in X \)

- **Variance of cluster i.**
  \[ \sigma_{v_i}^p = \frac{1}{n_i} \sum_{k=1}^{n_i} \left( x_k^p - v_i^p \right)^2 \]

- **Average scattering for clusters.**
  \[ \text{Scat}(c) = \frac{1}{c} \sum_{i=1}^{c} \left\| \sigma(v_i) \right\| \]

- **Total separation between clusters.**
  \[ \text{Dis}(c) = \frac{D_{\text{max}}}{D_{\text{min}}} \sum_{k=1}^{c} \left( \sum_{z=1}^{c} \left\| v_k - v_z \right\| \right)^{-1} \]

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
SD Index Definition

\[ SD(c) = a \times \text{Scat}(c) + \text{Dis}(c) \]

\( a = \text{Dis}(c_{\text{max}}) \), where \( c_{\text{max}} \) is the maximum number of input clusters.

- SD proposes an optimal number of clusters almost irrespectively of \( c_{\text{max}} \).
- SD handle properly convex clusters. The same applies to all the aforementioned indices.
The objective is to seek clustering schemes where most of the vectors of the dataset exhibit high degree of membership in one cluster.

A **fuzzy clustering** is defined by

- a matrix $U=\{u_{ij}\}$, where $u_{ij}$ denotes the degree of membership of the vector $x_i$ in the $j$ cluster.
- a set of the cluster representatives.

To evaluate clustering schemes

- we define **validity index**, $q$, and
- we plot the $q$ versus **number of clusters**.

If $q$ exhibits a trend with respect to the **number of clusters**, we seek a **significant knee** of decrease (or increase) in the plot of $q$. 

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Validity Indices involving only the membership values

**Partition coefficient**

\[
PC = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{nc} u_{ij}^2
\]

The PC index values range in \([1/nc, 1]\), where \(nc\) is the number of clusters.

- **If** \(PC \to 1\) **indicates** crisp clustering
- **If** \(PC = 1/nc\) **indicates** the fuzzy clustering or there is no clustering tendency in the considered dataset or the clustering algorithm failed to reveal it.

**Partition entropy**

\[
PE = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{nc} u_{ij} \cdot \log_a(u_{ij})
\]

The index is computed for \(nc > 1\) and \(PE \in [0, \log_a(nc)]\).

- **If** \(PE \to 0\), **indicates** crisp clustering
- **If** \(PE \to \log_a(nc)\) **indicates** absence of any clustering structure in the dataset or inability of the algorithm to extract it.
Validity Indices involving only the membership values

Drawbacks

- their monotonous dependency on the number of clusters. Thus, we seek significant knees of increase (for PC) or decrease (for PE) in plot of the indices versus the number of clusters,

- their sensitivity to the fuzzifier, $m$. More specifically, as $m \to 1$ the indices give the same values for all values of $n_c$. On the other hand when $m \to \infty$, both PC and PE exhibit significant knee at $n_c=2$,

- the lack of direct connection to the geometry of the data [Dave96], since they do not use the data itself.
Indices involving the membership values and the dataset.

Xie-Beni index

Let a fuzzy partition of the data set \( X=\{x_j; j=1,\ldots, n\} \) with \( v_i(i=1,\ldots, nc) \) the centers of each cluster and \( u_{ij} \) the membership of data point \( j \) belonging to cluster \( i \).

Compactness of cluster \( i \)

\[ \pi=(\sigma_i/n_i). \]

\( n_i \) : the number of point in cluster belonging to cluster \( i \),

\( \sigma_i \) : variance of cluster \( i \)

Separation of the fuzzy partitions

\[ d_{\min} = \min \|v_i - v_j\| \]

\[ XB=\pi/N \cdot d_{\min} \]
Indices involving the membership values and the dataset.

- **Fukuyama-Sugeno index**

\[
FS_m = \sum_{i=1}^{N} \sum_{j=1}^{nc} u_{ij}^m \left( \|x_i - v_j\|_A^2 - \|v_j - v\|_A^2 \right)
\]

- \(v\): the mean vector of \(X\) and

- \(A\): an \(l \times l\) positive symmetric matrix, when \(A = I\), the above distance become the squared Euclidean distance.

- small values for \(FS_m\) → compact and well-separated clusters
Indices involving the membership values and the dataset.

- Fuzzy covariance matrix of the j-th cluster
  \[ \sum_j = \frac{\sum_{i=1}^{N} u_{ij}^m (x_i - v_j) (x_i - v_j)^T}{\sum_{i=1}^{N} u_{ij}^m} \]
  \[ V_j = |\Sigma_j|^{1/2} \]

- Fuzzy hyper volume of j-th cluster
  \[ FH = \sum_{j=1}^{nc} V_j \]

- Total fuzzy hyper volume
  \[ \sum \sum_{j=1}^{nc} V_j \]

- Average partition density
  \[ PA = \frac{1}{nc} \sum_{j=1}^{nc} S_j \]
  \[ S_j = \sum_{x \in X_j} u_{ij} \]

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
<table>
<thead>
<tr>
<th></th>
<th>DataSet1</th>
<th>DataSet2</th>
<th>DataSet3</th>
<th>DataSet4</th>
<th>Real_Data1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optimal number of clusters</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$RS, RMSST$</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>$DB$</td>
<td>6</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>$SD$</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Trends in Clustering

- Discovering and finding representatives of arbitrary shaped clusters.
- Non-point clustering.
- Handling uncertainty in the clustering process and visualization of results.
- Incremental clustering.
- Constraint-based clustering.

M. Halkidi, Y. Batistakis, M. Vazirgiannis, SSDBM, July 2001
Conclusions

★ The clustering aims at identifying and extract significant groups in underlying data.

★ Cluster analysis is a major tool in a number of applications in many fields of business and science.

★ The clustering algorithms can be classified into: i. partitional, ii. hierchical, iii. density-based, iv. grid-based, iv. fuzzy clustering

★ The cluster validity is an important issue of clustering. It is related to the inherent features of the data set under concern.

★ A survey of the most known validity criteria is presented. They are classified in three categories: external, internal, and relative.

★ An experimental evaluation of some representative validity indices are presented
Thank you for your attention!