Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- Major Clustering Methods
- Outlier Analysis

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
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

Examples of Clustering Applications

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

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Requirements of Clustering in Data Mining

- Scalability
  - Sampling a large data set gives biased results
  - Need highly scalable clustering algorithms
- Ability to deal with different types of attributes
  - Not only interval-based numerical data, but also
    - Binary, categorical, ordinal, or a combination of these
- Discovery of clusters with arbitrary shape
  - Not just spherical clusters based on Euclidean or Manhattan distance

- Minimal requirements for domain knowledge to determine input parameters
  - # of desired clusters, etc.
- Able to deal with noise and outliers
  - Clusters should not be of poor quality
- Insensitive to order of input records
  - Same clusters should be generated

Cluster Analysis

- What is Cluster Analysis?
- **Types of Data in Cluster Analysis**
- Major Clustering Methods
- Outlier Analysis

Data Structures

- Types of data that occur often in cluster analysis
- Preprocessing the data
- # of objects – *n*
- Data structures of main memory based algorithms:
  - Data matrix (object-by-variable structure)
    - n objects (e.g., persons)
    - p variables (measurements or attributes)
      - age, weight, height, etc
    - n – by – p matrix

- Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix (object-by-object structure)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Measure the Quality of Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric: $d(i, j)$
- There is a separate "quality" function that measures the "goodness" of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define "similar enough" or "good enough"
  - the answer is typically highly subjective.

Type of data in clustering analysis

- Interval-scaled variables:
- Binary variables:
- Nominal, ordinal, and ratio variables:
- Variables of mixed types:

Interval-valued variables

- Variables that have continuous measurements
  - E.g., weight, height, temperature, …
- How does the units affect the clustering?
  - Meters to inches
  - Kgs to lbs
    - Will change the cluster behavior
    - Smaller units will lead to a larger range for that variable
- Independent of choice of measurement units
  - Standardize the data
  - Give all variables equal weight

Interval-valued variables

- Standardize data
  - Calculate the **mean absolute deviation, s_f, of attribute f**

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf}).$$

  - Calculate the standardized measurement (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation is more robust to outliers than using standard deviation

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan (or city block) distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

Similarity and Dissimilarity Between Objects (Cont.)

- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$ *(symmetric)*
    - $d(i,j) \leq d(i,k) + d(k,j)$ *(triangular inequality)*
- If each variable is assigned a weight

$$d(i,j) = \sqrt{w_1(|x_{i_1} - x_{j_1}|^2 + w_2|x_{i_2} - x_{j_2}|^2 + \ldots + w_p|x_{i_p} - x_{j_p}|^2)}$$

Binary Variables

- A contingency table for binary data

		Object j		
		1	0	*sum*
	1	q	r	$q+r$
Object i	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- All binary variables have equal weight
- q - number of variables that equal 1 for both objects i and j
- Similarly, r, s, and t
- Total number of variables is p = q + r + s + t

Binary Variables

- Dissimilarity between objects i and j
  - Simple matching coefficient (if the binary variable is *symmetric*):

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

    - Symmetric: both states are equally valuable
      - Example: gender (male or female, either can be coded as 0 or 1)

- Jaccard coefficient (if the binary variable is *asymmetric*):

$$d(i, j) = \frac{r + s}{q + r + s}$$

  - Example: positive and negative outcomes of a disease test
    - The agreement of two 1s (a positive match) is more significant that that of two 0s (a negative match)
    - Number of negative matches, t, is considered unimportant and thus ignored

Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	1	0	1	0	0	0
Mary	F	1	0	1	0	1	0
Jim	M	1	1	0	0	0	0

  - gender is a *symmetric* attribute
  - the remaining attributes are *asymmetric* binary

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

- Highest dissimilarity value
  - Jim and Mary
  - Hence, unlikely to have a similar disease

Nominal Variables

- A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables, dissimilarity:

$$d(i,j) = \frac{p-m}{p}$$

- Method 2: use a large number of binary variables
  - creating a new binary variable for each of the *M* nominal states

Ordinal Variables

- An ordinal variable can be discrete or continuous
- order is important, e.g., rank of professor
- Can be treated like interval-scaled
  - replacing x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- **Major Clustering Methods**
- Outlier Analysis

Major Clustering Approaches

- Partitioning algorithms: Construct various partitions and then evaluate them by some criterion
- Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Partitioning Algorithms: Basic Concept

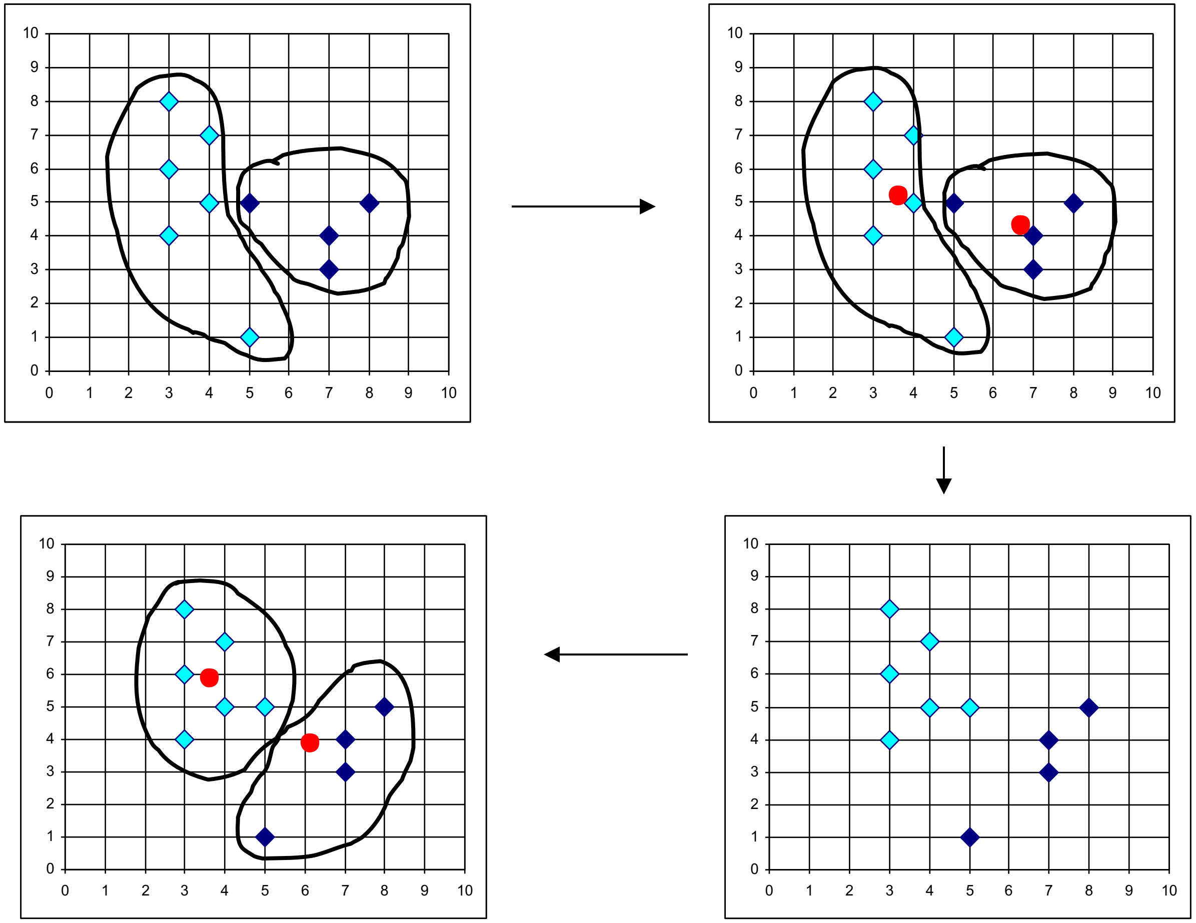
- <u>Partitioning method:</u> Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means* and *k-medoids* algorithms
  - <u>*k-means*</u> : Each cluster is represented by the center of the cluster
  - <u>*k-medoids*</u> or PAM (Partition around medoids) : Each cluster is represented by one of the objects in the cluster

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in 4 steps:
  1. Partition objects into *k* nonempty subsets
  2. Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
  3. Assign each object to the cluster with the nearest seed point.
  4. Go back to Step 2, stop when no more new assignment.

The *K-Means* Clustering Method

- Example

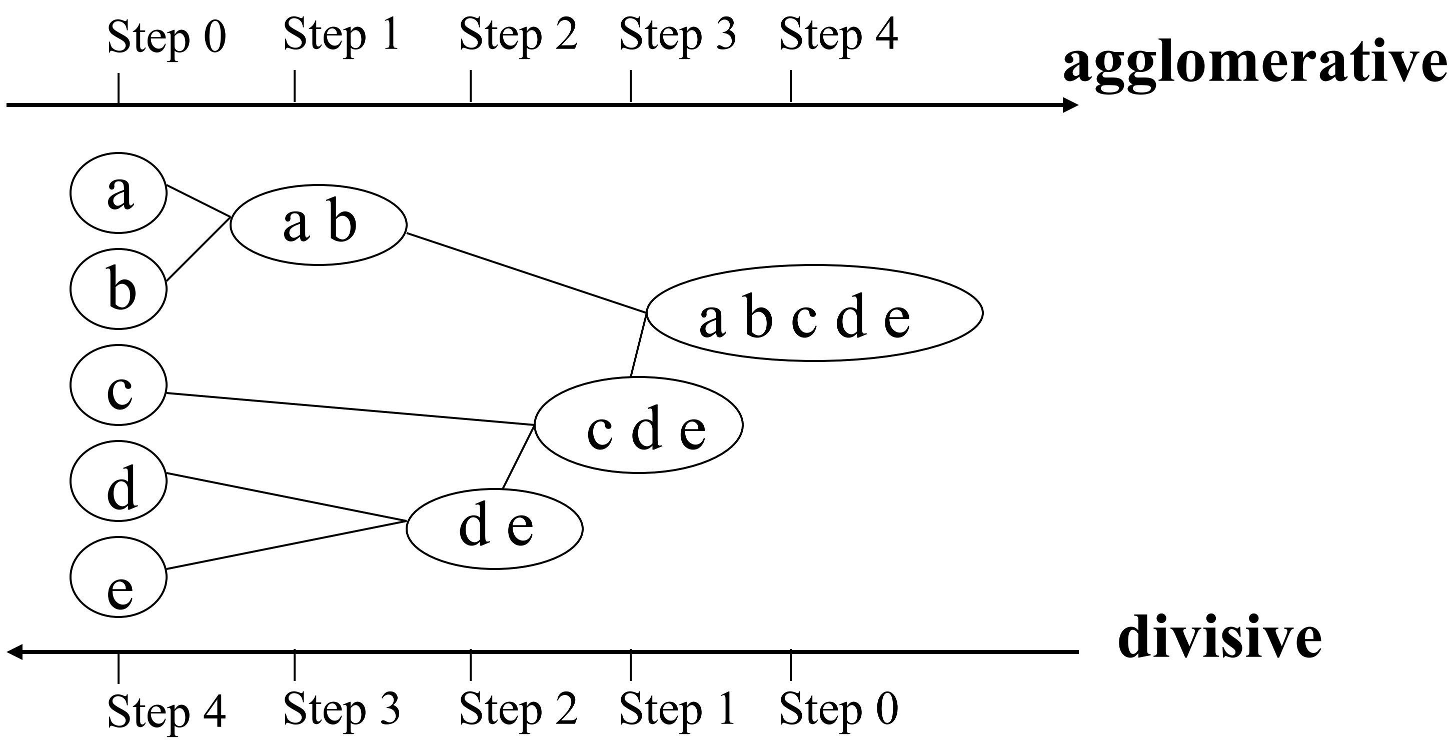


The *K-Medoids* Clustering Method

- Drawback of k-means
  - Sensitive to outliers
- Not the mean value as the reference point in a cluster
- Find *representative* objects, called medoids, in clusters
  - The most centrally located object in a cluster
  - Minimize the sum of the dissimilarities between each object and its medoid

Hierarchical Clustering

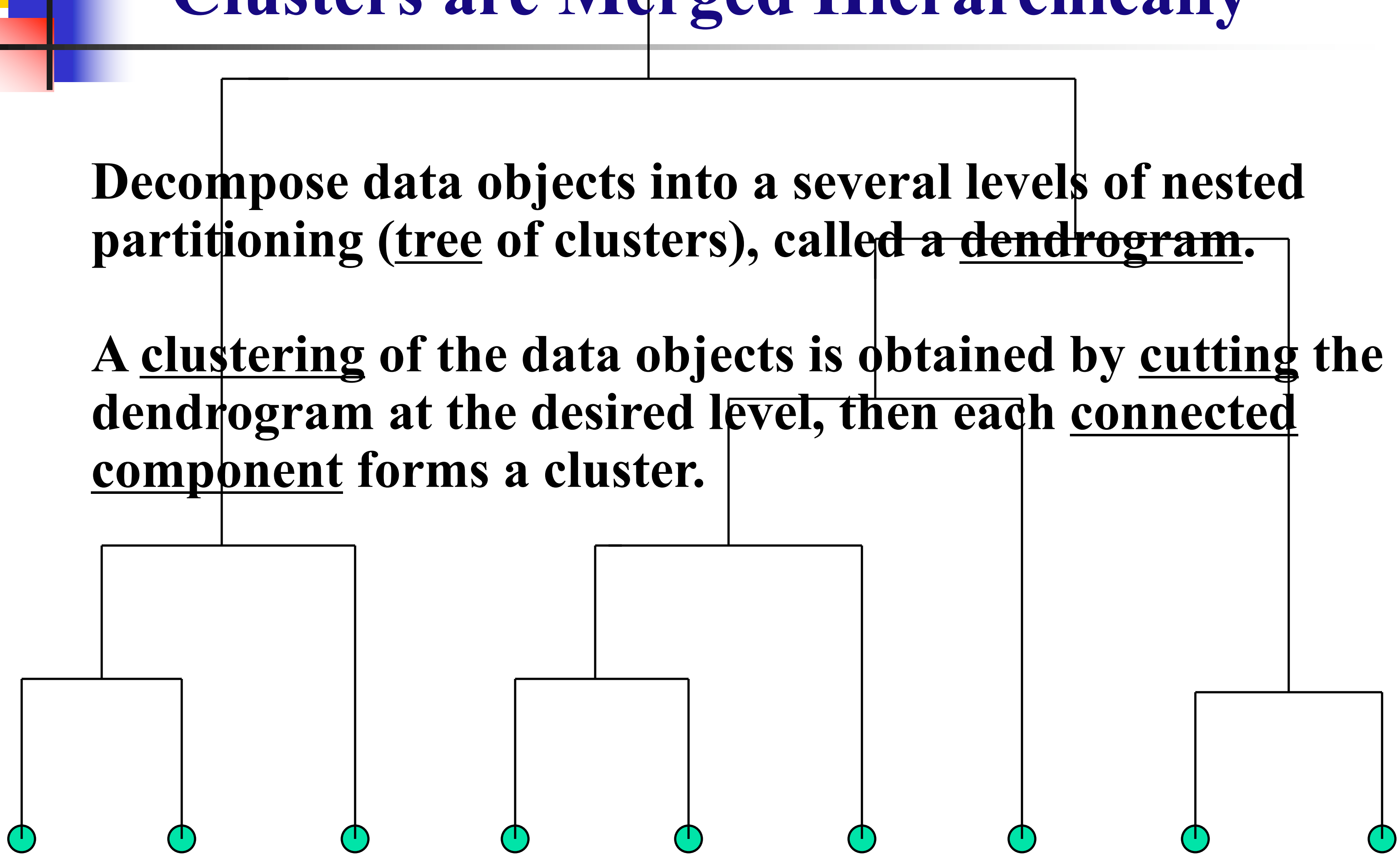
- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition



A *Dendrogram* Shows How the Clusters are Merged Hierarchically

Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.



Cluster Analysis

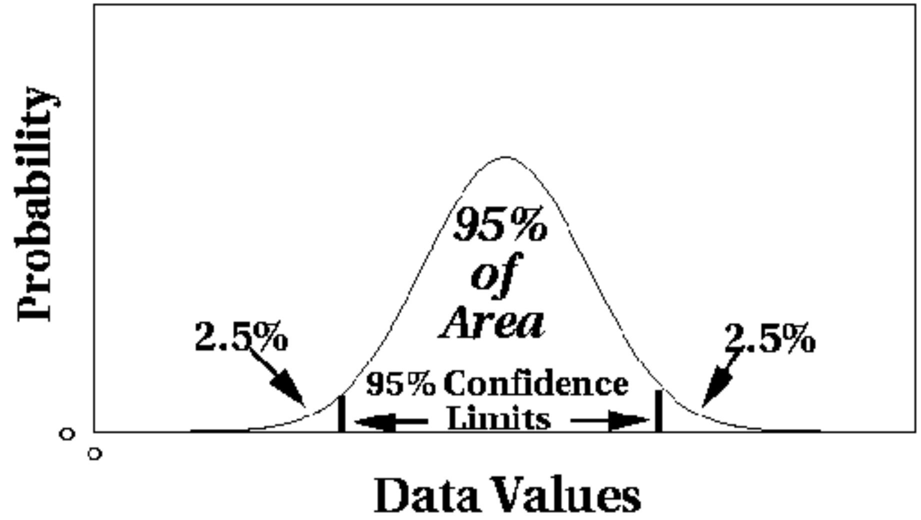
- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- Major Clustering Methods
- **Outlier Analysis**

What Is Outlier Discovery?

- What are outliers?
  - The set of objects are considerably dissimilar from the remainder of the data
    - Example: Sports: Michael Jordon, Wayne Gretzky, ...
      Salary of CEO
  - Should we minimize the influence of outliers or eliminate them?
    - One person's noise could be another person's signal
- Outlier Mining
  - Given *n* data objects, and *k*, the expected number of outliers
    - Find the top *k* objects that are considerably dissimilar with respect to the remaining data

- Applications:
  - Credit card fraud detection
  - Telecom fraud detection
    - Unusual usage
  - Customer segmentation
    - Extremely low or extremely high incomes
  - Medical analysis
    - Unusual responses to various medical treatments

Outlier Discovery: Statistical Approach



- Assume a model underlying distribution that generates data set (e.g. normal distribution)
- Use discordancy tests depending on
  - data distribution
  - distribution parameter (e.g., mean, variance)
  - number of expected outliers
- Drawbacks
  - most tests are for single attribute
  - In many cases, data distribution may not be known