Cluster Analysis - Abridged

CS4881 Aritificial Intelligence

Jay Urbain, PhD

Credits:

Tom Mitchell, Machine Learning

Jiawei Han and Micheline Kamber, Data Mining

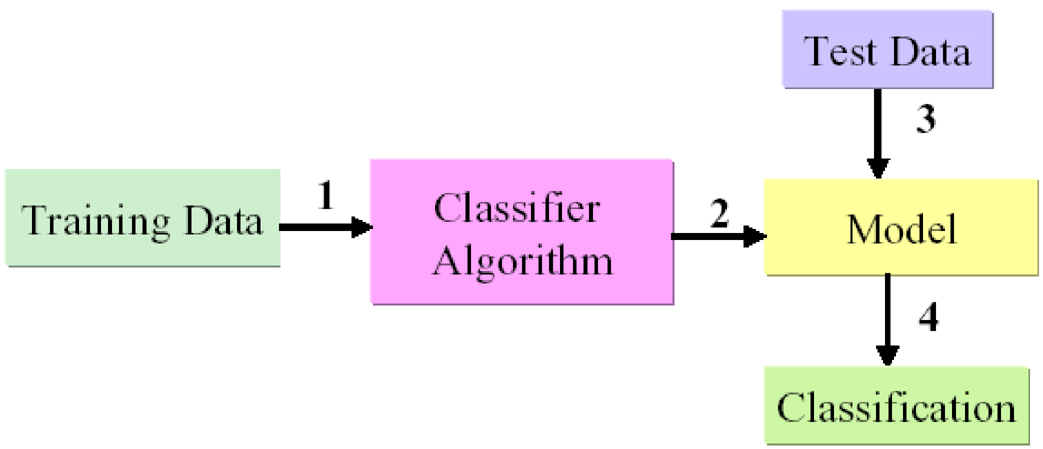
Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Summary

Supervised Learning

Supervised Learning

- Learn by example from training data with a class label
- Create model by running algorithm on training data
- Identify a class label for the incoming new data



Supervised Learning Algorithms

Supervised Learning Algorithms

- Naive Bayes
- Neural Networks
- Decision Trees
- Support Vector Machine
- Bayes Nets

Unsupervised Learning Algorithms

There are many machine learning situations in which class labels are not available, so *unsupervised* methods are needed.

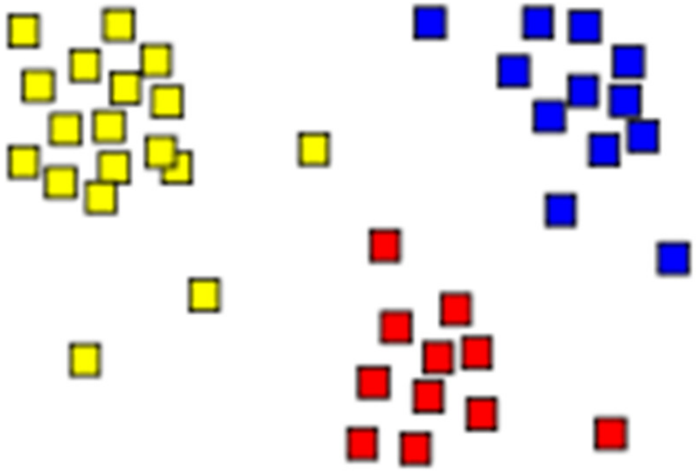
Unsupervised Learning Algorithms

- Clustering – *many*
- Topic Models
- Collaborative Filtering
- Association Rules - *Apriori*

Clustering

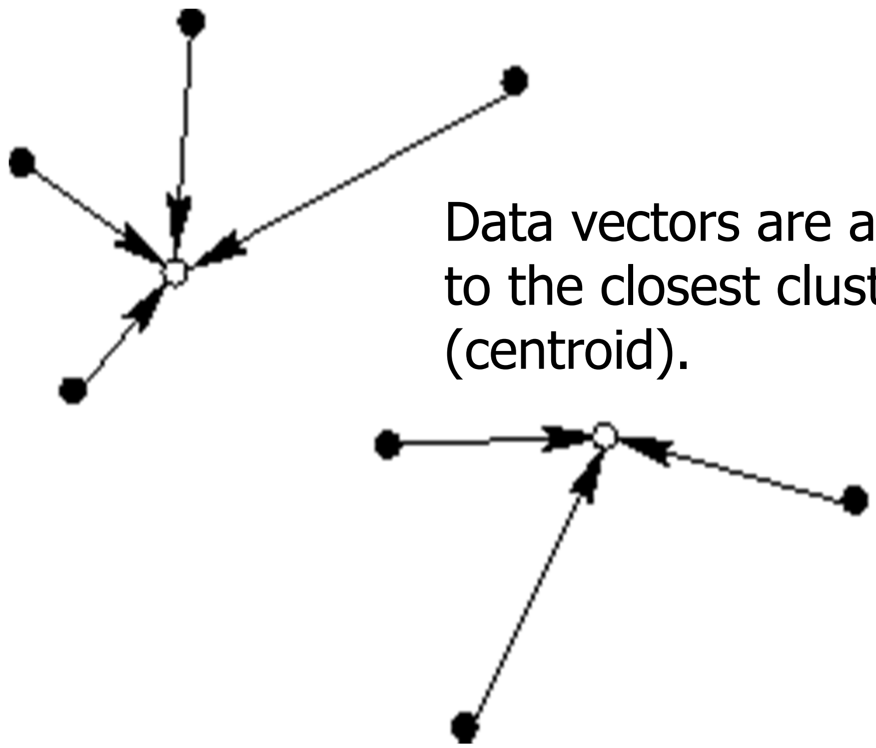
Clustering: assigning a set of objects into groups (**clusters**) so that the objects in the same cluster are more ***similar*** (with respect to some features) to each other than to those in other clusters.

Cluster analysis itself is not one specific algorithm (there are many), but the general task to be solved.



Clustering

Objects are *clustered* into groups that reflect distinct regions of the decision (feature) space.



Data vectors are assigned to the closest cluster center (centroid).

One of the first clustering applications?

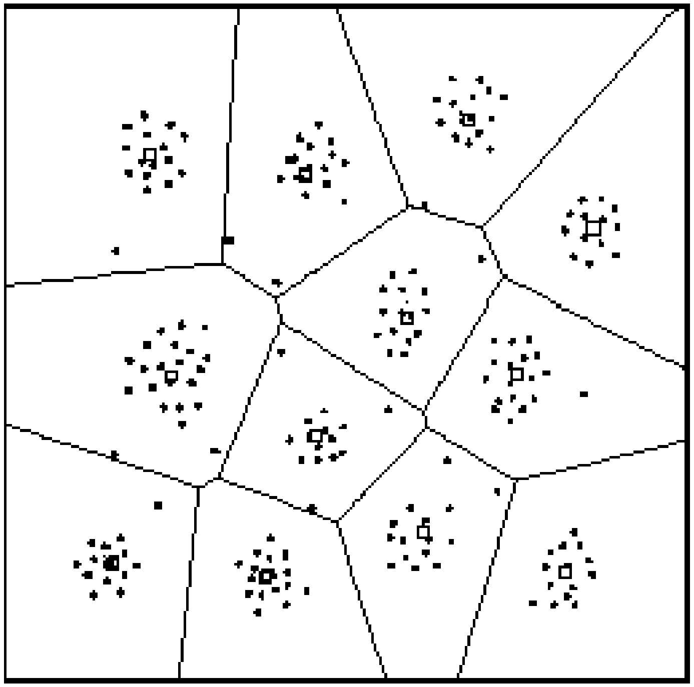
During a cholera outbreak in London in 1854, John Snow used a special map to plot the cases of the disease that were reported.

A key observation, after the creation of the map, was the close association between the density of disease cases and a single well located at a central street.

After this, the well pump was removed putting an end to the epidemic.

Cluster Analysis Defined

- 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

- Exploratory data mining technique

General Applications

- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

- General Applications
  - Co-expressed genes
  - NLP Semantics
  - Document clustering for IR
  - Pattern Recognition
  - Spatial Data Analysis
  - Image Processing
  - Market research
  - Cluster Weblog data discover search groups

Specific Applications

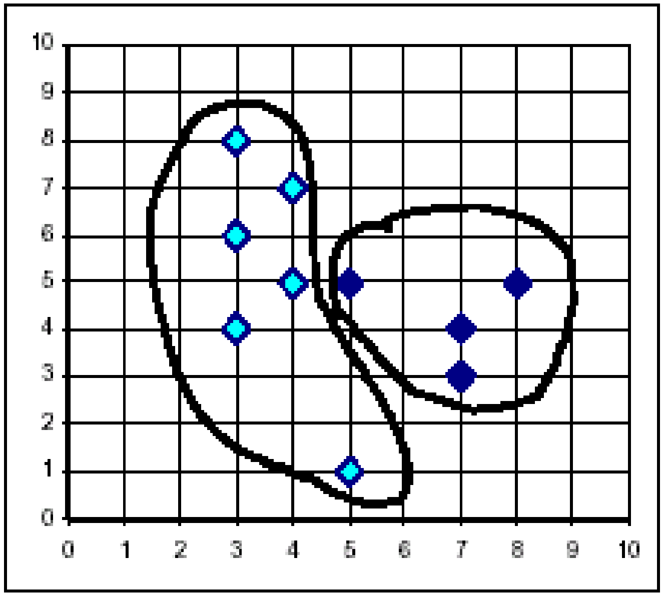
- Social Networking: Collaborative filtering.
- Marketing: Discover distinct groups for targeted marketing programs.
- Land use: Identification of areas of similar land use.
- Insurance: Identifying groups with a high claim cost.
- City-planning: Identifying groups of houses according to their house type, value, and geographical location.
- Earth-quake studies: Epicenters should be clustered along continent faults.
- Text mining: Identify frequently co-occuring terms for concept identification.

What Is Good Clustering?

- A *good clustering* method will produce high quality clusters with:
  - high intra-class similarity
  - low inter-class similarity

- Dependent on method used, data/domain, and implementation.

- Quality measured by its ability to discover some or all of the hidden patterns.



Measuring Distance for Object Assignment

- Similarity is expressed in terms of a distance function, which is typically a metric: $d(i, j)$.
- There is a separate "quality" function that measures the "goodness" of a cluster.
- Distance functions are different for *interval-scaled*, *boolean*, *nominal/categorical*, *ordinal*, and *ratio* variables.
- Hard to define "similar enough" or "good enough" – its subjective.

Data types must be normalized

- Nominal
  - Categorical data. Use labels/categories. E.g., rocks as igneous, sedimentary and metamorphic; eye color, zip codes.
- Ordinal
  - Rankings. E.g., movie rating on a 1-4 star scale, grades, height in {tall, medium, short}, places in a race.
- Interval
  - Measurable on interval scales. Celsius 1/100.
  - Any difference between levels of an attribute can be multiplied by any real number to exceed or equal another difference.
- Ratio
  - Measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind.
  - Comparison of one value to another. E.g., temperature in Kelvin, %.

Similarity/Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects

- Some popular ones include: Manhattan distance

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

Similarity and Dissimilarity Between Objects (Cont.)

- Euclidean distance:

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

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- Also one can use weighted distance, Pearson product correlation, or other dissimilarity measures.

Similarity Between Objects - Cosine

- *Cosine distance*
- *Normalized cross-product of 2 vectors.*
- *Popular in information retrieval.*

$$sc(i,j) = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Normalizing Interval Data

- Standardize data
  - Calculate the *mean absolute deviation*:

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

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

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

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

- Note: Using mean absolute deviation is more robust than using standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Binary Variables

- A contingency table for binary data

		Object j		
		1	0	sum
	1	a	b	$a+b$
Object i	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Simple matching coefficient (invariant, if the binary variable is *symmetric*):

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- Jaccard coefficient (*noninvariant* if the binary variable is *asymmetric*) disregard negative matches, i.e., positive and negative matches do not contain symmetric information:

$$d(i,j) = \frac{b+c}{a+b+c}$$

Binary Variables - Jaccard Index

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute ***similarities*** using the following quantities
  - M01 = the number of attributes where p was 0 and q was 1
  - M10 = the number of attributes where p was 1 and q was 0
  - M00 = the number of attributes where p was 0 and q was 0
  - M11 = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

  = (M11 + M00) / (M01 + M10 + M11 + M00)

- Jaccard = number of 11 matches / number of not-both-zero attributes values = (M11) / (M01 + M10 + M11)

SMC vs. Jaccard Index

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

M01 = 2 (number of attributes where p was 0 and q was 1)

M10 = 1 (number of attributes where p was 1 and q was 0)

M00 = 7 (number of attributes where p was 0 and q was 0)

M11 = 0 (number of attributes where p was 1 and q was 1)

SMC = (M11 + M00)/(M01 + M10 + M11 + M00)

= (0+7) / (2+1+0+7) = 0.7

J = (M11) / (M01 + M10 + M11)

= 0 / (2 + 1 + 0) = 0

Nominal (Categorical) 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 (typical)
  - *m*: # of matches, *p*: total # of variables

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

- 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
- 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

Ratio-Scaled Variables

- Ratio-scaled variable:
- Methods:
  - treat them like interval-scaled variables (linear data)
  - apply logarithmic transformation (exponential data: Ae^{Bt} or Ae^{-Bt})
  - treat them as continuous ordinal data, treat their rank as interval-scaled.

Variables of Mixed Types

- A database may contain all the six types of variables
  - symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio.
- One may use a weighted formula to combine their effects.

$$d(i,j) = \frac{\Sigma_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\Sigma_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - f is binary or nominal:

    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$
  - f is interval-based: use the normalized distance
  - f is ordinal or ratio-scaled
    - compute ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
    - and treat z_{if} as interval-scaled

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
- Statistical/Model Based: A model is hypothesized for each of the clusters. The idea is to find the best fit of that model to each other. E.g., Expectation Maximum, Topic Model
- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure

Partitioning Algorithms: Basic Concept

- Partitioning method: 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
    - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster (centroid)
    - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster (most representative)

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in 4 steps:
  1. Partition objects (records) 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 or *convergence*.

K-Means Psuedo Code

```
Randomly assign samples to each cluster
for each Cj // for each cluster
        Cj.sum <=Sum(Vi in j'th cluster)
        Cj.count <= Sum(1 if Vi in j'th cluster)
        Cj.centroid <= Cj.sum/Cj.count
Do
    for each Vi       // object
        for each Cj // for each cluster
            Dij <= Euclidian (Vi,Cj) // distance between object and cluster centroid
            if( Dij < Vi.d )
                Vi.d = Dij
                Vi.centroid = Cj
        Assign Vi to centroid Cx with smallest distance

    for each Cj // for each cluster
        Cj.sum <=Sum(Vi in j'th cluster)
        Cj.count <= Sum(1 if Vi in j'th cluster)
        Cj.centroid <= Cj.sum/Cj.count

While (Diff <THRESHOLD)
```

Comments on the *K-Means* Method

- <u>Strength</u>
  - *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, k, $t << n$.
  - Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *annealing* and *genetic algorithms.*
- <u>Weakness</u>
  - Applicable only when *mean* is defined. Not well suited for categorical data.
  - Need to specify *k,* the *number* of clusters, in advance.
  - Unable to handle noisy data and *outliers.*
  - Not suitable to discover clusters with *non-convex shapes.*

Variations of the *K-Means* Method

- A few variants of the *k-means* which differ in
  - Selection of the initial *k*
  - Distance calculations
    - single-linkage clustering (the minimum of object distances)
    - complete linkage clustering (the maximum of object distances)
    - UPGMA("Unweighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering).
  - Strategies to calculate cluster means

K-Modes Method

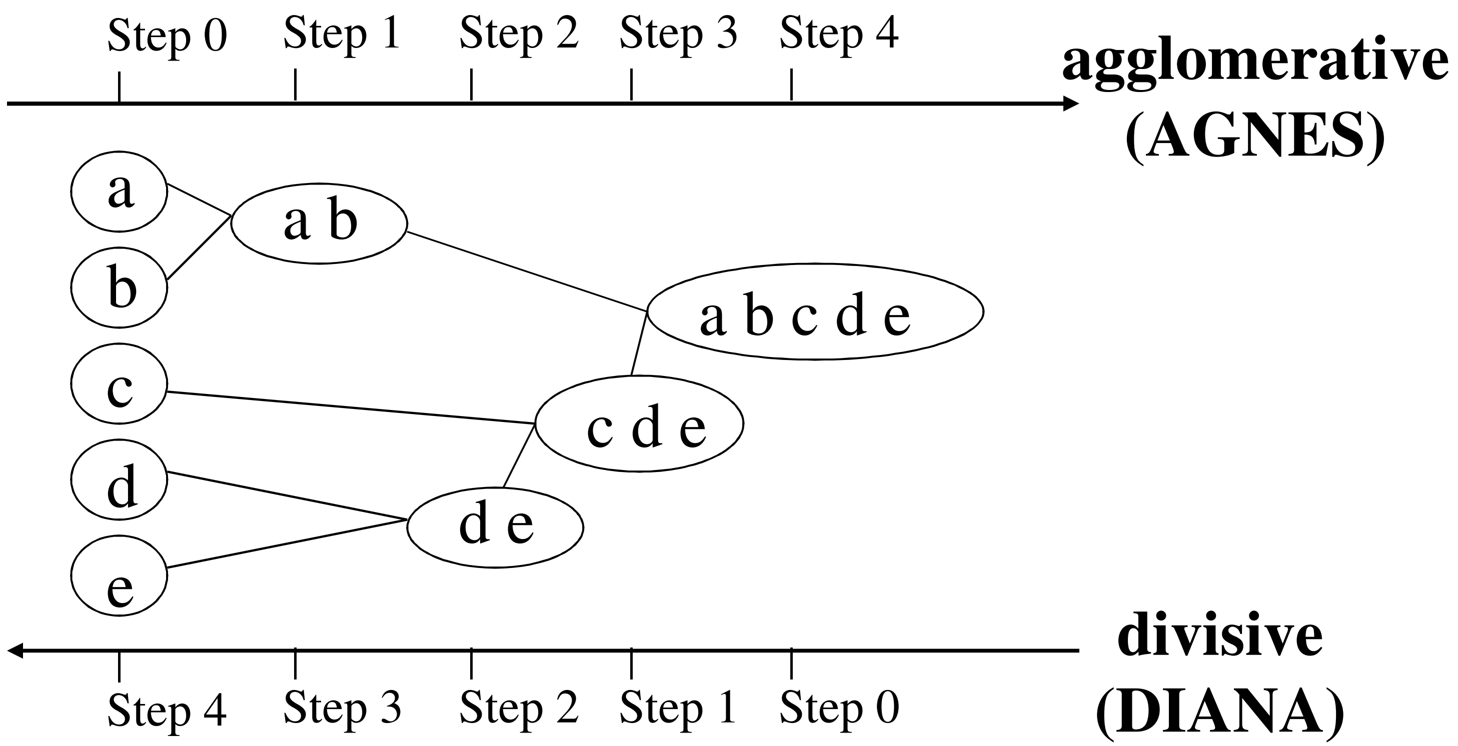
- Handling categorical data: *k-modes* (Huang'98)
  - Replacing means of clusters with modes
  - Using new dissimilarity measures to deal with categorical objects
  - Using a frequency-based method to update modes of clusters
  - A mixture of categorical and numerical data: *k-prototype* method

The *K-Medoids* Clustering Method

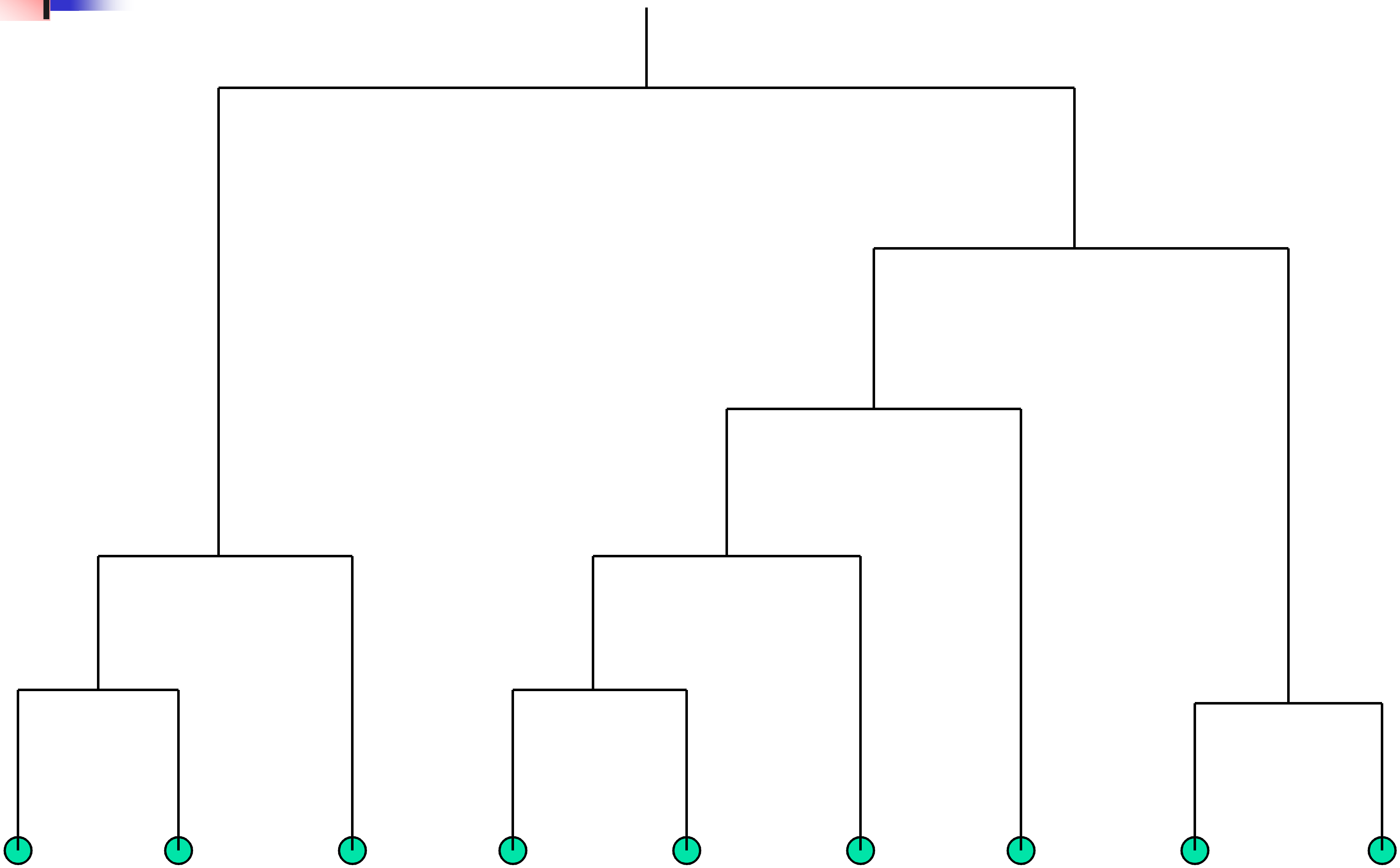
- Find *representative* objects, called medoids, in clusters
- *PAM* (Partitioning Around Medoids, 1987)
  - starts from an initial set of *medoids* and iteratively replaces one of the *medoids* by one of the non-medoids if it improves the total distance of the resulting clustering,
  - *PAM* works effectively for small data sets, but does not scale well for large data sets,
- *CLARA* (Kaufmann & Rousseeuw, 1990)
- *CLARANS* (Ng & Han, 1994): Randomized sampling
- Focusing + spatial data structure (Ester et al., 1995)

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

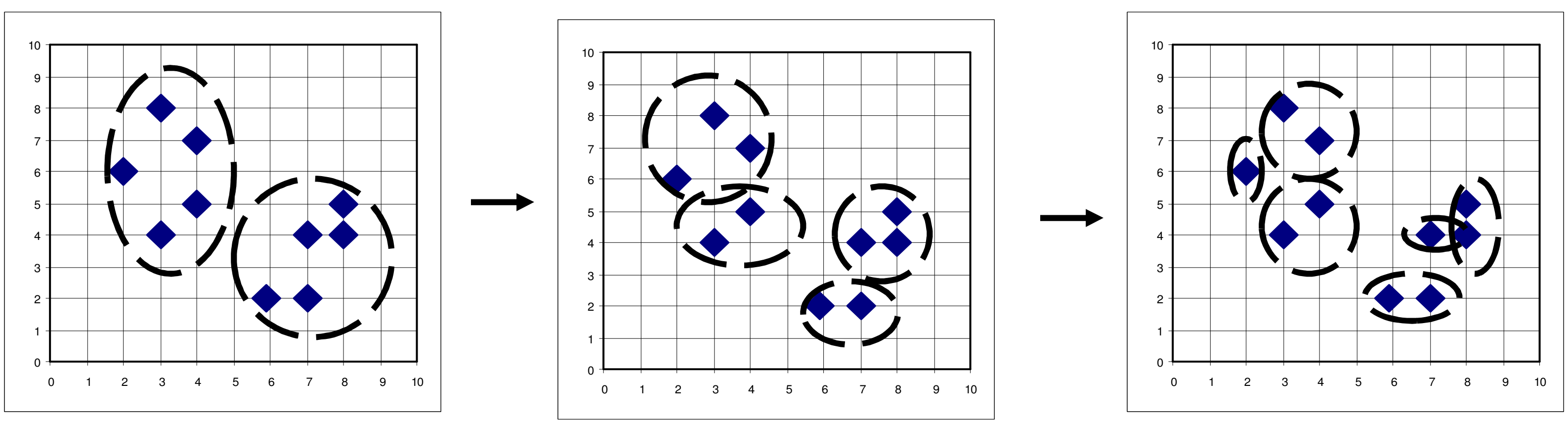


A *Dendrogram* Shows How the Clusters are Merged Hierarchically



DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own



More on Hierarchical Clustering Methods

- Major weakness of agglomerative clustering methods
  - <u>does not scale</u> well: time complexity of at least $O(n^2)$, where n is the number of total objects
  - can never undo what was done previously
- Demo clustering of search results!

Summary

- Cluster analysis groups objects based on their similarity and has wide applications.
- Measure of similarity can be computed for various types of data.
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods.
- Outlier detection and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches.
- There are still lots of research issues on cluster analysis, such as constraint-based clustering, statistical methods.

- Note: Demo Topic Models if time!

References (1)

- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD'98
- M. R. Anderberg. Cluster Analysis for Applications. Academic Press, 1973.
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure, SIGMOD'99.
- P. Arabie, L. J. Hubert, and G. De Soete. Clustering and Classification. World Scietific, 1996
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. KDD'96.
- M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. SSD'95.
- D. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2:139-172, 1987.
- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. In Proc. VLDB'98.
- S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. SIGMOD'98.
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.

References (2)

- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB’98.
- G. J. McLachlan and K.E. Bkasford. Mixture Models: Inference and Applications to Clustering. John Wiley and Sons, 1988.
- P. Michaud. Clustering techniques. Future Generation Computer systems, 13, 1997.
- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.
- E. Schikuta. Grid clustering: An efficient hierarchical clustering method for very large data sets. Proc. 1996 Int. Conf. on Pattern Recognition, 101-105.
- G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. VLDB’98.
- W. Wang, Yang, R. Muntz, STING: A Statistical Information grid Approach to Spatial Data Mining, VLDB’97.
- T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH : an efficient data clustering method for very large databases. SIGMOD'96.

Data Structures

- Data matrix
  - (two modes)

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

- Dissimilarity matrix
  - (one mode)

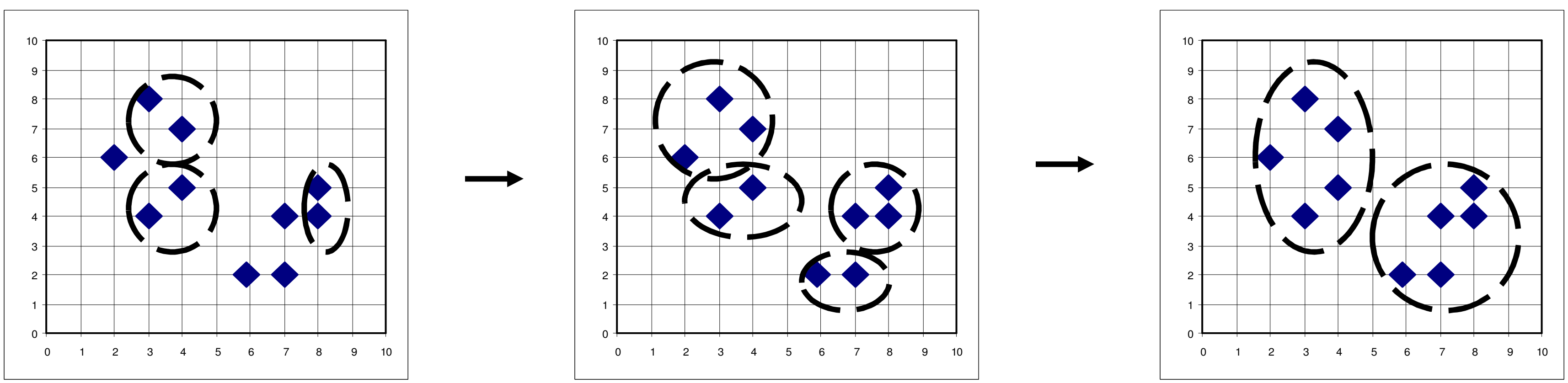
$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Data (Stevens, Science 1946) (skip)

Scale Type	Permissible Statistics	Admissible Scale Transformation	Mathematical structure
nominal (also denoted as categorical)	mode, chi square	One to One (equality (=))	standard set structure (unordered)
ordinal	median, percentile	Monotonic increasing (order (<))	totally ordered set
interval	mean, standard deviation, correlation, regression, analysis of variance	Positive linear (affine)	affine line
ratio	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation, logarithms	Positive similarities (multiplication)	field

AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster



Binary Variables (skip)

- Jaccard index is a statistic used for comparing the similarity and diversity of sample sets. Jaccard similarity coefficient:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

- Jaccard distance coefficient:

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

The *K-Means* Clustering Method

- Example