



DWDM : Unit VI

Cluster Analysis Introduction

UNIT – VI: Cluster Analysis Introduction: Introduction to machine learning, Types of Data in Cluster Analysis, A Categorization of Major Clustering Methods, Partitioning Methods, Density-Based Methods, Grid-Based Methods, Model-Based Clustering Methods, Outlier Analysis.

Unit VI. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods (Beyond Curriculum)
- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Outlier Analysis
- Summary

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

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted mktg. programs
- Land use: Identification of areas of similar land use in an earth observation db.
- 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
- Earth-quake studies: 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
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Chapter 8. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
 - Partitioning Methods
 - Hierarchical Methods
 - Density-Based Methods
 - Grid-Based Methods
 - Model-Based Clustering Methods
- Outlier Analysis
- Summary

Data Structures

- Data matrix-
Object by Variable
 - (two modes)

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

- Dissimilarity matrix
 - (one mode)

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

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

- Standardize data

- Calculate the mean absolute deviation (MAD):

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

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

where

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

- Calculate the standardized measurement (z-score)

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

- Using mean absolute deviation is more robust 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*:

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

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

- If $q = 1$, d is Manhattan distance

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

Similarity and Dissimilarity Between Objects (Cont.)

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

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

- 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, parametric Pearson product moment correlation, or other dissimilarity measures.

Binary Variables

- A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	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):

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

Dissimilarity between Binary Variables

■ Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- gender is a symmetric attribute
- the remaining attributes are asymmetric binary
- let the values Y and P be set to 1, and the value N be set to 0

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

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

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

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

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

- 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 $r_{if} \in \{1, \dots, M_f\}$
 - 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: a positive measurement on a nonlinear scale, approximately at exponential scale, such as Ae^{Bt} or Ae^{-Bt}
- Methods:
 - treat them like interval-scaled variables — *not a good choice! (why?)*
 - apply logarithmic transformation
$$y_{if} = \log(x_{if})$$
 - 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.

- f is binary or nominal:
$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$
 - $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$ or missing values, or $d_{ij}^{(f)} = 1$ n.w.
- f is ordinal or ratio-scaled
 - compute ranks r_{if} and
 - and treat z_{if} as interval-scaled
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Chapter 8. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Outlier Analysis
- Summary

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

Chapter 8. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
 - Partitioning Methods
 - Hierarchical Methods
 - Density-Based Methods
 - Grid-Based Methods
 - Model-Based Clustering Methods
 - Outlier Analysis
- Summary

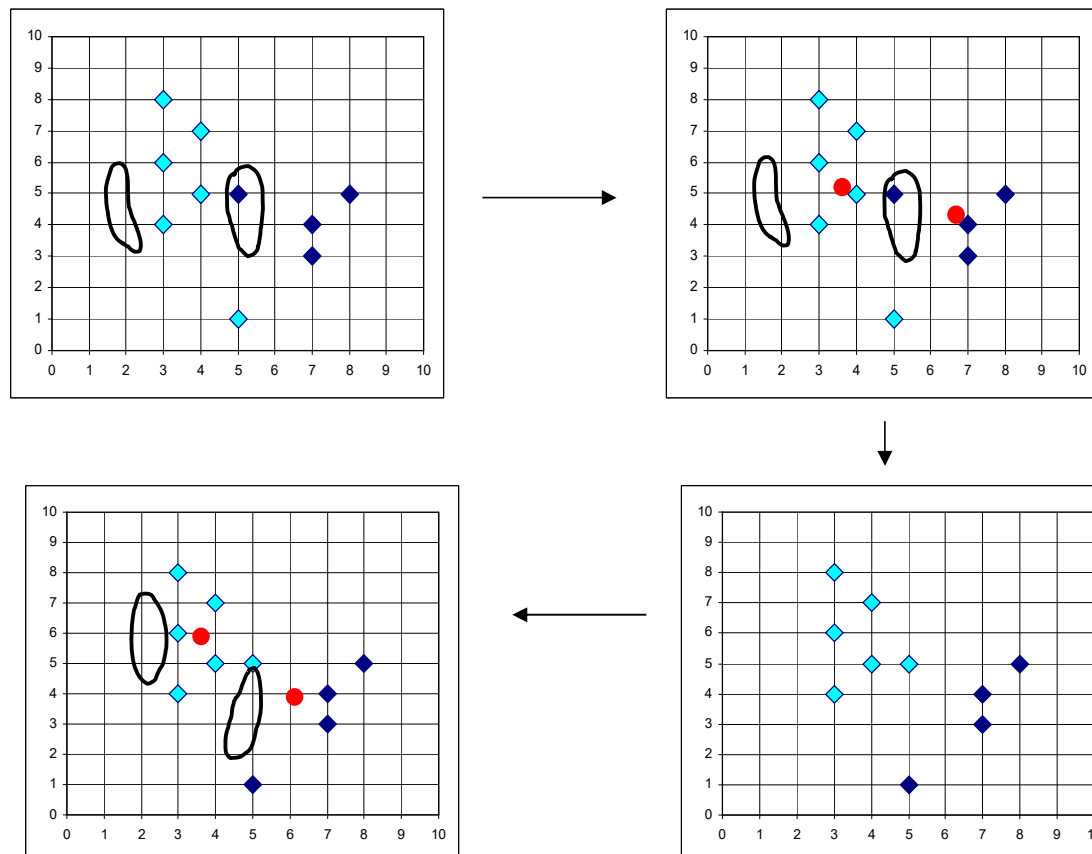
Partitioning Algorithms: Basic Concept

- 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
 - k -medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): 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



Comments on the *K-Means* Method

- Strength

- *Relatively efficient: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$.*
- *Often terminates at a local optimum. The global optimum may be found using techniques such as: deterministic annealing and genetic algorithms*

- Weakness

- *Applicable only when mean is defined, then what about 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* means
 - Dissimilarity calculations
 - Strategies to calculate cluster means
- 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

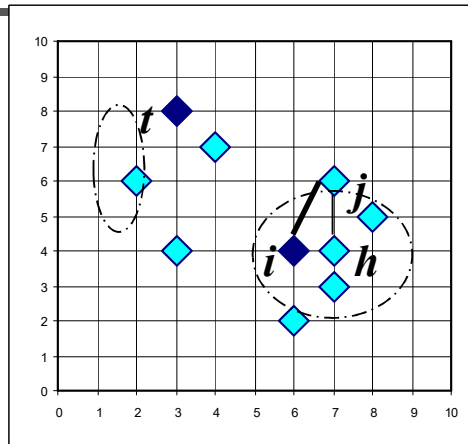
The *K-Medoids* Clustering Method

- 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)

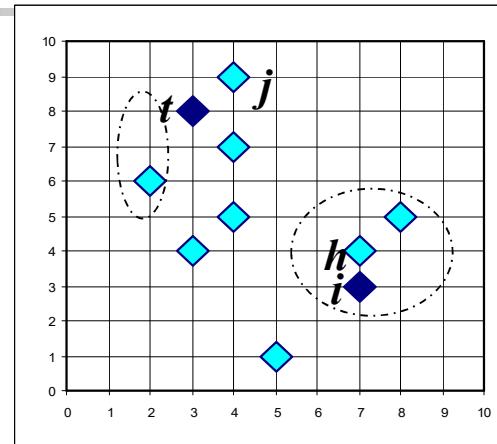
PAM (Partitioning Around Medoids) (1987)

- PAM (Kaufman and Rousseeuw, 1987), built in Splus
- Use real object to represent the cluster
 1. Select k representative objects arbitrarily
 2. For each pair of non-selected object h and selected object i , calculate the total swapping cost TC_{ih}
 3. For each pair of i and h ,
 - a. If $TC_{ih} < 0$, i is replaced by h
 - b. Then assign each non-selected object to the most similar representative object
 4. repeat steps 2-3 until there is no change

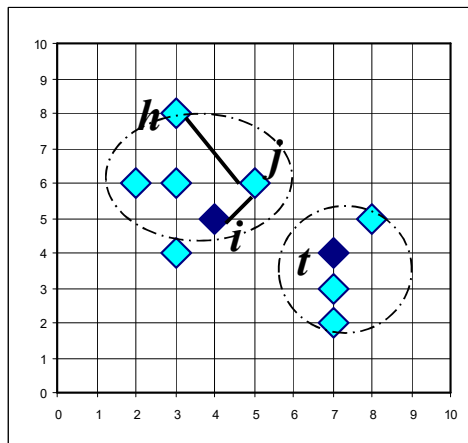
PAM Clustering: Total swapping cost $TC_{ih} = \sum_j C_{jih}$



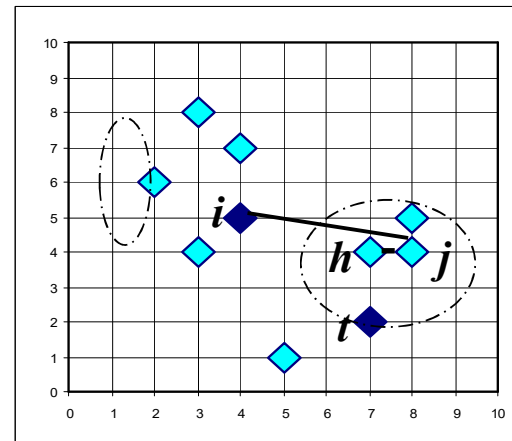
$$C_{jih} = d(j, h) - d(j, i)$$



$$C_{jih} = 0$$

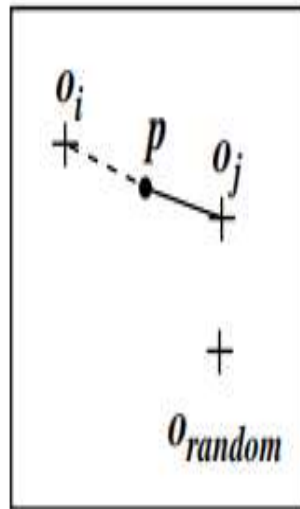


$$C_{jih} = d(j, t) - d(j, i)$$

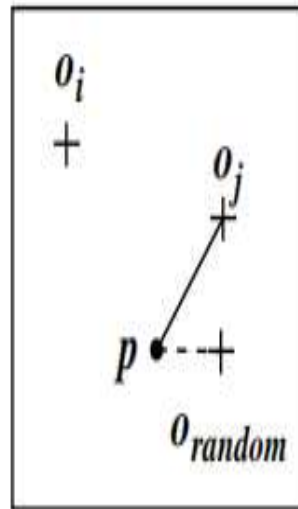


$$C_{jih} = d(j, h) - d(j, t)$$

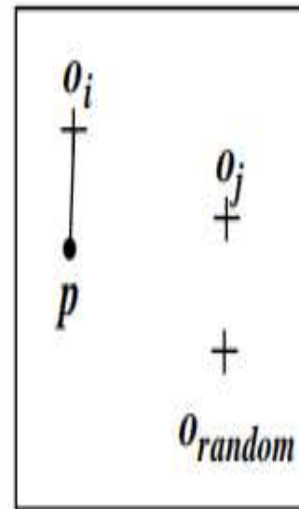
4 cases of K- Medoids Clustering



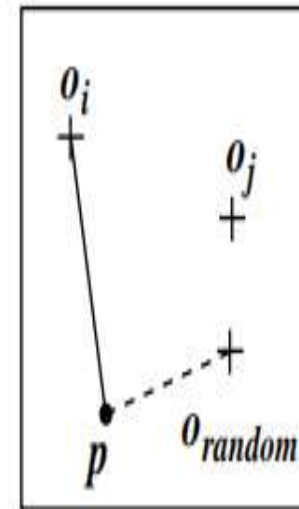
(a) Reassigned
to o_i



(b) Reassigned
to o_{random}



(c) No change



(d) Reassigned
to o_{random}

- Data object
- + Cluster center
- Before swapping
- - - After swapping

CLARA (Clustering Large Applications) (1990)

- CLARA (Kaufmann and Rousseeuw in 1990)
 - Built in statistical analysis packages, such as S+
- It draws *multiple samples* of the data set, applies *PAM* on each sample, and gives the best clustering as the output
- Strength: deals with larger data sets than *PAM*
- Weakness:
 - 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) (1994)

- CLARANS (A Clustering Algorithm based on Randomized Search) (Ng and Han'94)
- CLARANS draws sample of neighbors dynamically
- The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of k medoids
- If the local optimum is found, CLARANS starts with new randomly selected node in search for a new local optimum
- It is more efficient and scalable than both *PAM* and *CLARA*
- Focusing techniques and spatial access structures may further improve its performance (Ester et al.'95)

Chapter 8. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Outlier Analysis
- Summary

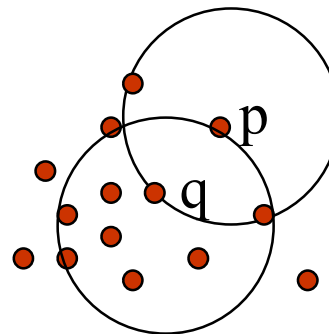
Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
 - Discover clusters of arbitrary shape
 - Handle noise
 - One scan
 - Need density parameters as termination condition
- Several interesting studies:
 - DBSCAN: Ester, et al. (KDD'96)
 - OPTICS: Ankerst, et al (SIGMOD'99).
 - DENCLUE: Hinneburg & D. Keim (KDD'98)
 - CLIQUE: Agrawal, et al. (SIGMOD'98)

Density-Based Clustering: Background

- Two parameters:
 - **Eps**. Maximum radius of the neighbourhood
 - **MinPts**. Minimum number of points in an Eps-neighbourhood of that point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid \text{dist}(p,q) \leq Eps\}$
- Directly density-reachable: A point p is directly density-reachable from a point q wrt. **Eps**, **MinPts** if
 - 1) p belongs to $N_{Eps}(q)$
 - 2) core point condition:

$$|N_{Eps}(q)| \geq \text{MinPts}$$



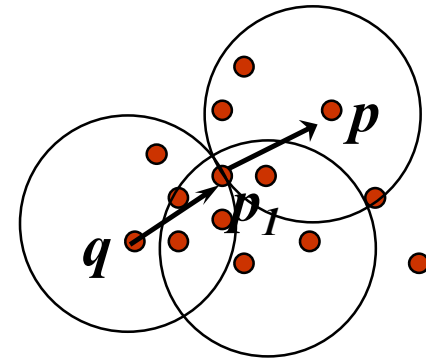
MinPts = 5

Eps = 1 cm

Density-Based Clustering: Background (II)

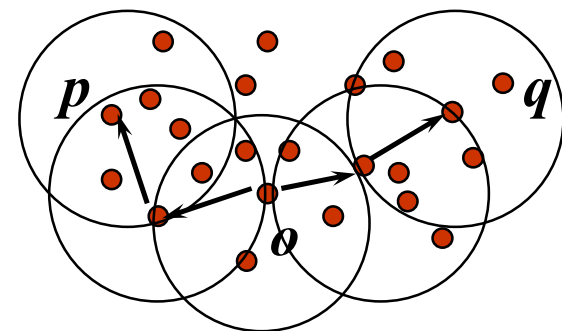
- Density-reachable:

- A point p is density-reachable from a point q wrt. Eps , $MinPts$ if there is a chain of points p_1, \dots, p_n , $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i



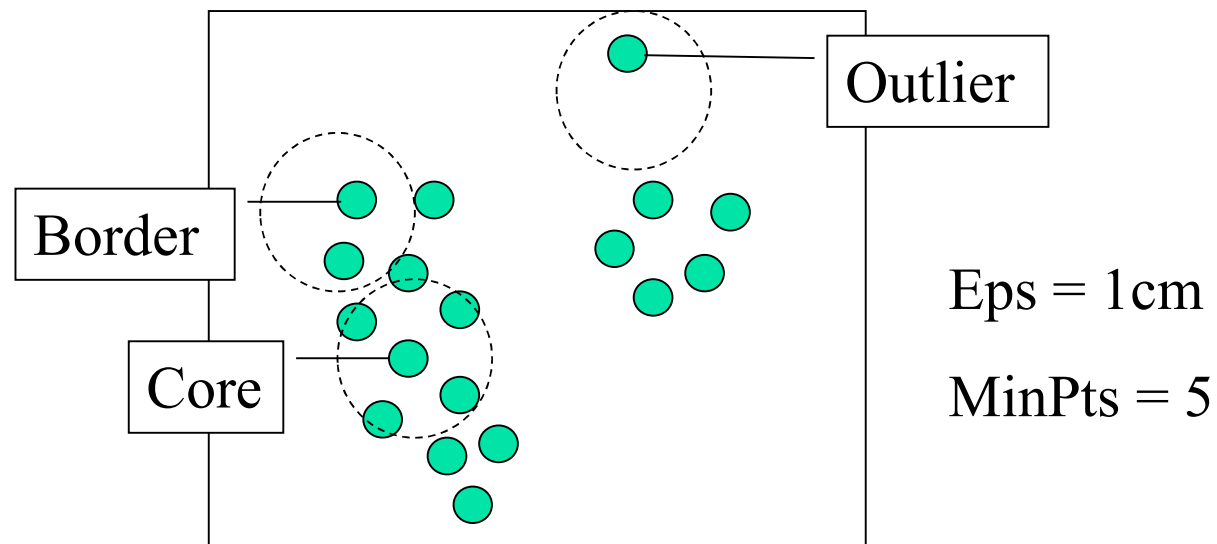
- Density-connected

- A point p is density-connected to a point q wrt. Eps , $MinPts$ if there is a point o such that both, p and q are density-reachable from o wrt. Eps and $MinPts$.



DBSCAN: Density Based Spatial Clustering of Applications with Noise

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



DBSCAN: The Algorithm

- Arbitrary select a point p
- Retrieve all points density-reachable from p wrt ***Eps*** and ***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.

OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure
 - Produces a special order of the database wrt its density-based clustering structure
 - This cluster-ordering contains info equiv to the density-based clusterings corresponding to a broad range of parameter settings
 - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
 - Can be represented graphically or using visualization techniques

OPTICS: Some Extension from DBSCAN

- Index-based:
 - k = number of dimensions
 - $N = 20$
 - $p = 75\%$
 - $M = N(1-p) = 5$

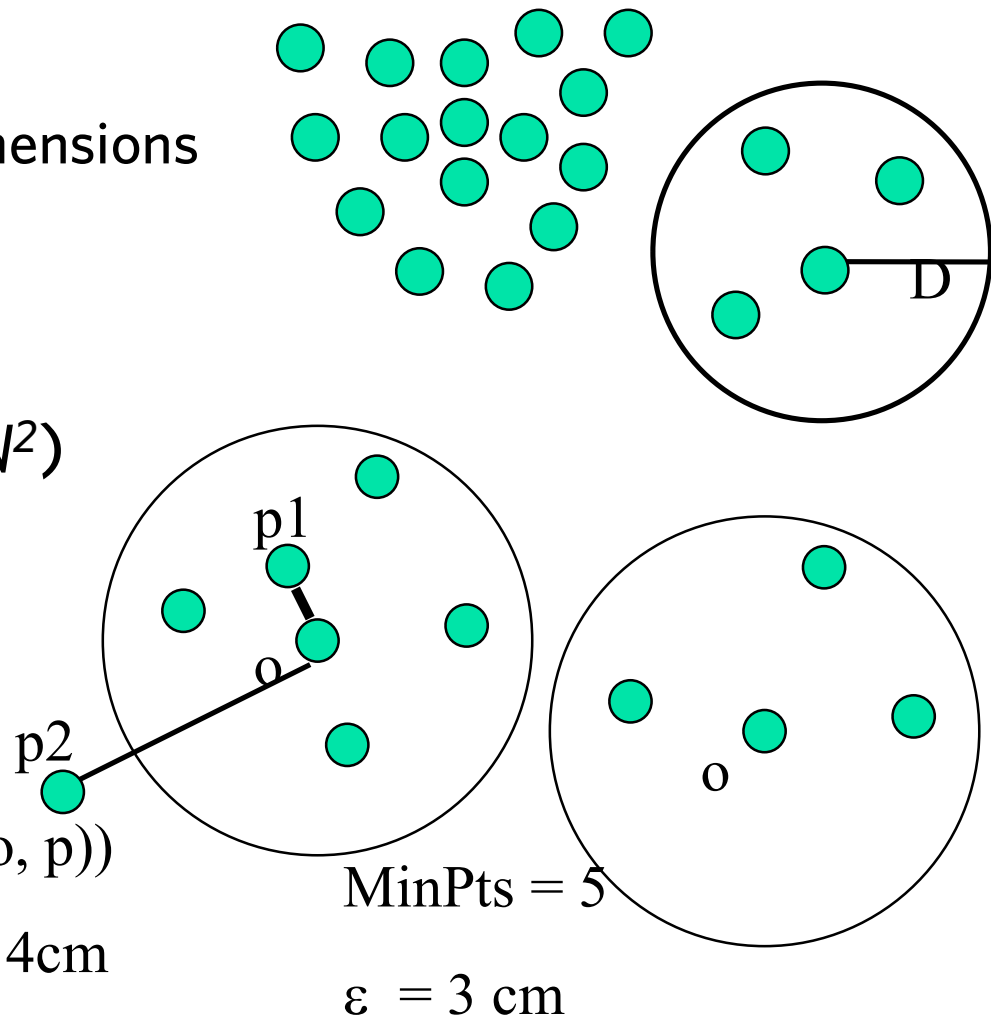
■ Complexity: $O(kN^2)$

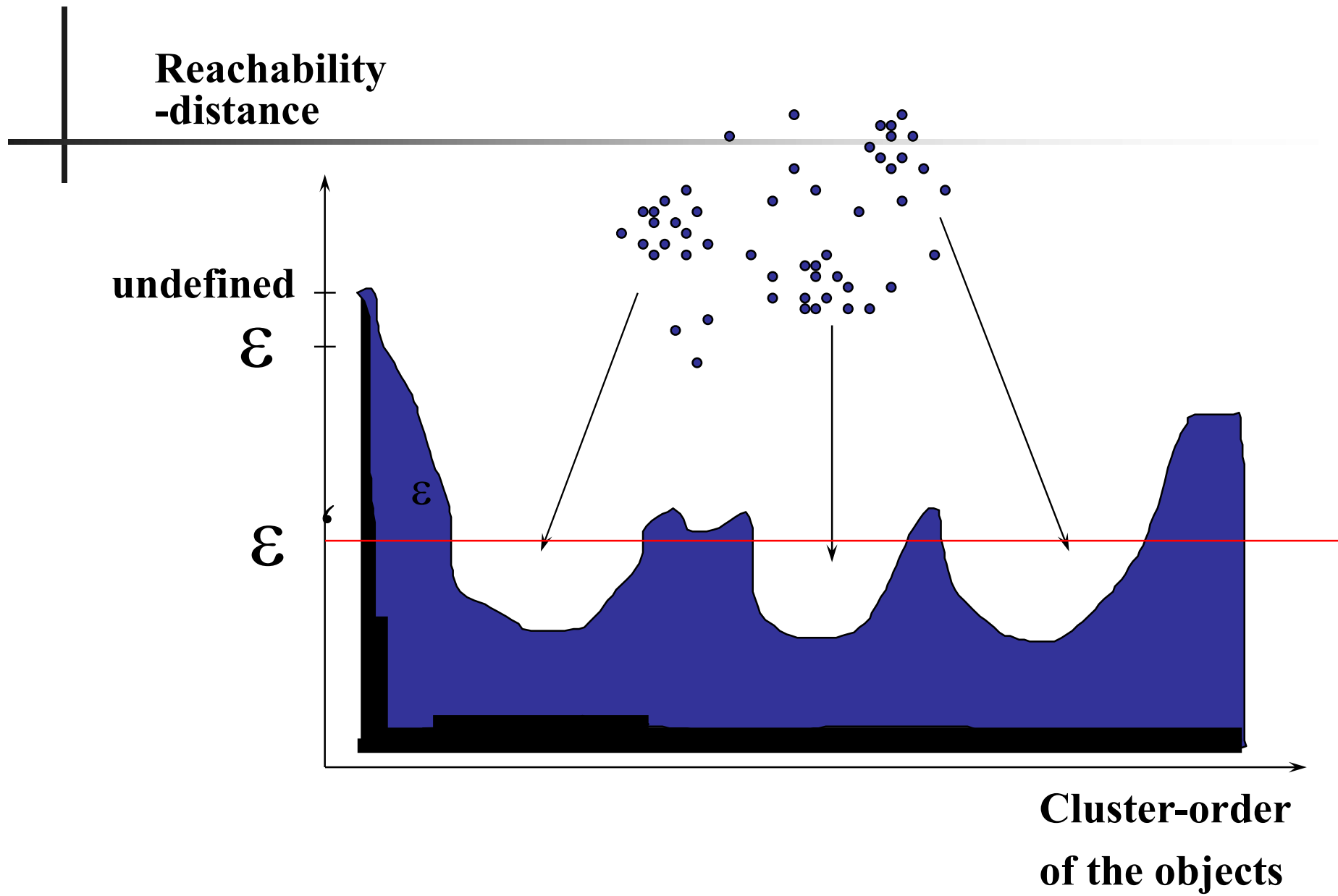
- Core Distance

- Reachability Distance

$\text{Max}(\text{core-distance}(o), d(o, p))$

$r(p1, o) = 2.8\text{cm}$. $r(p2, o) = 4\text{cm}$





DENCLUE: using density functions

- DENSity-based CLUstEring by Hinneburg & Keim (KDD'98)
- Major features
 - Solid mathematical foundation
 - Good for data sets with large amounts of noise
 - Allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets
 - Significant faster than existing algorithm (faster than DBSCAN by a factor of up to 45)
 - But needs a large number of parameters

Denclue: Technical Essence

- Uses grid cells but only keeps information about grid cells that do actually contain data points and manages these cells in a tree-based access structure.
- 1. **Influence function:** describes the impact of a data point within its neighborhood.
- 2. Overall density of the data space can be calculated as the sum of the influence function of all data points.
- 3. Clusters can be determined mathematically by identifying density attractors.
- Density attractors are local maximal of the overall density function.

Gradient: The steepness of a slope

■ Example

$$f_{\text{Gaussian}}(x, y) = e^{-\frac{d(x, y)^2}{2\sigma^2}}$$

$$f_{\text{Gaussian}}^D(x) = \sum_{i=1}^N e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$

$$\nabla f_{\text{Gaussian}}^D(x, x_i) = \sum_{i=1}^N (x_i - x) \cdot e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$

Chapter 8. Cluster Analysis

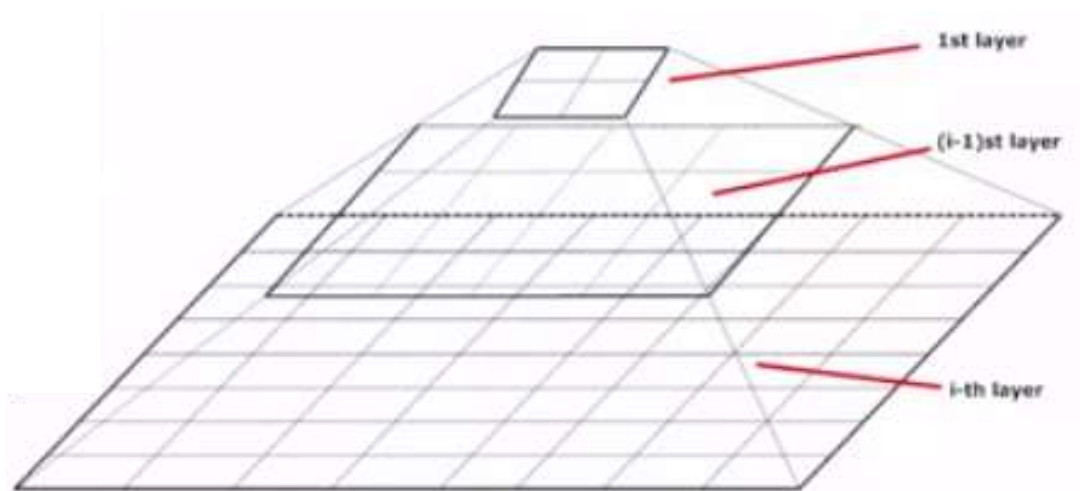
- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
 - Partitioning Methods
 - Hierarchical Methods
 - Density-Based Methods
 - Grid-Based Methods
 - Model-Based Clustering Methods
- Outlier Analysis
- Summary

Grid-Based Clustering Method

- Using multi-resolution grid data structure
- Several interesting methods
 - **STING** (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
 - **WaveCluster** by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
 - A multi-resolution clustering approach using wavelet method
 - **CLIQUE**: Agrawal, et al. (SIGMOD'98)

STING: A Statistical Information Grid Approach

- The spatial area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution



STING: A Statistical Information Grid Approach (2)

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level
- Statistical info of each cell is calculated and stored beforehand and is used to answer queries
- Parameters of higher level cells can be easily calculated from parameters of lower level cell
 - *count, mean, s, min, max*
 - type of distribution—normal, *uniform*, etc.
- Use a top-down approach to answer spatial data queries
- Start from a pre-selected layer—typically with a small number of cells
- For each cell in the current level compute the confidence interval

STING: A Statistical Information Grid Approach (3)

- Remove the irrelevant cells from further consideration
- When finish examining the current layer, proceed to the next lower level
- Repeat this process until the bottom layer is reached
- Advantages:
 - Query-independent, easy to parallelize, incremental update
 - $O(K)$, where K is the number of grid cells at the lowest level
- Disadvantages:
 - All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

WaveCluster (1998)

- A multi-resolution clustering approach which applies wavelet transform to the feature space
 - A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-band.
- Both grid-based and density-based
- Input parameters:
 - # of grid cells for each dimension
 - the wavelet, and the # of applications of wavelet transform.

WaveCluster (1998)

- Why is wavelet transformation useful for clustering
 - Unsupervised clustering
 - It uses hat-shape filters to emphasize region where points cluster, but simultaneously to suppress weaker information in their boundary
 - Effective removal of outliers
 - Multi-resolution
 - Cost efficiency
- Major features:
 - Complexity $O(N)$
 - Detect arbitrary shaped clusters at different scales
 - Not sensitive to noise, not sensitive to input order
 - Only applicable to low dimensional data

CLIQUE (Clustering In QUEst)

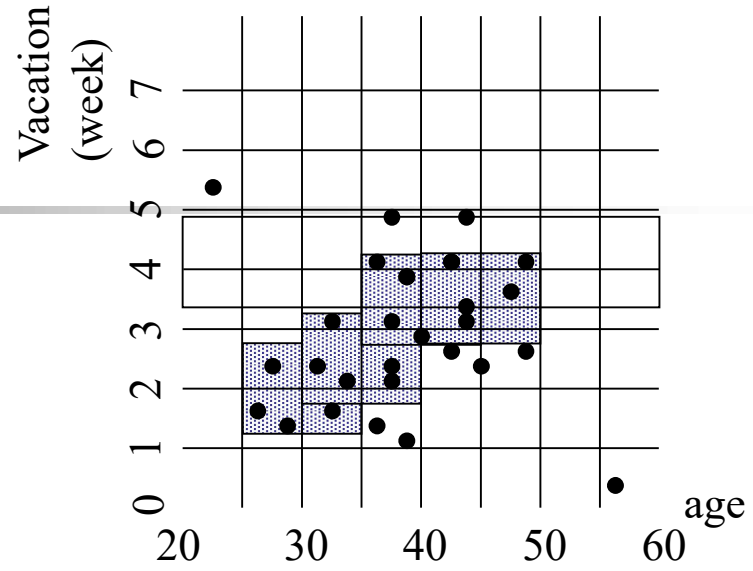
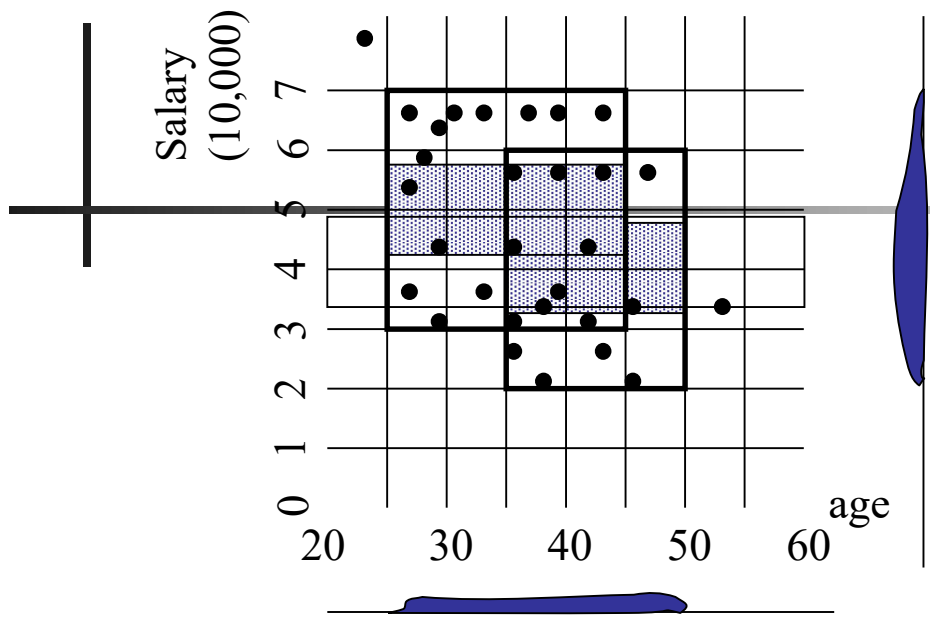
- Automatically identifying subspaces of a high dimensional data space that allow better clustering than original space
- CLIQUE can be considered as both **density-based and grid-based**
 1. It partitions each dimension into the same number of equal length interval
 2. It partitions an m-dimensional data space into non-overlapping rectangular units
 3. A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
 4. A cluster is a maximal set of connected dense units within a subspace

CLIQUE: Grid-Based Subspace Clustering

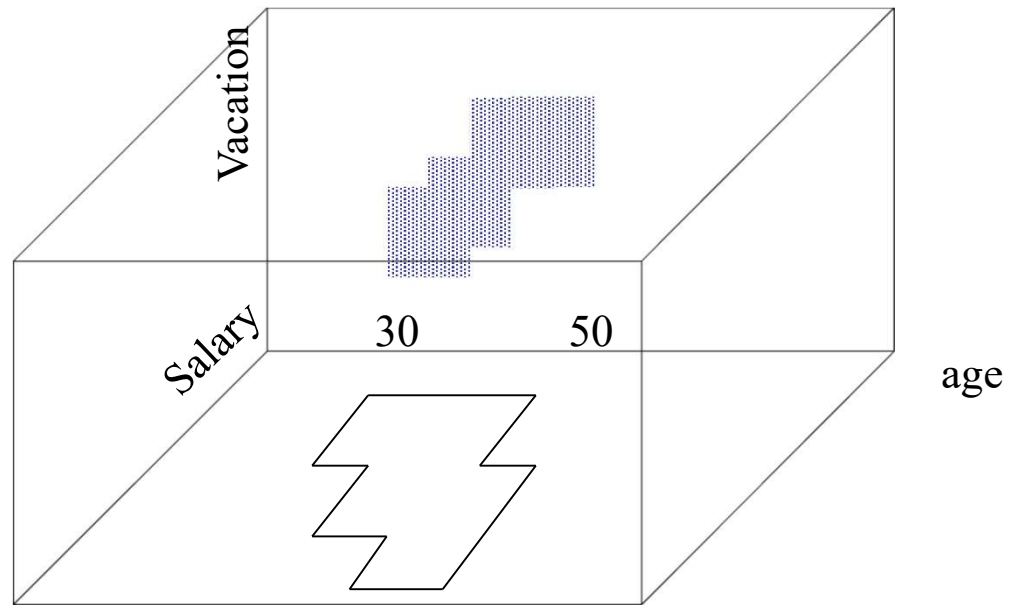
- ❑ CLIQUE (Clustering In QUEst) (Agrawal, Gehrke, Gunopulos, Raghavan: SIGMOD'98)
- ❑ CLIQUE is a **density-based** and **grid-based** **subspace clustering** algorithm
 - ❑ **Grid-based**: It discretizes the data space through a grid and estimates the density by counting the number of points in a grid cell
 - ❑ **Density-based**: A cluster is a maximal set of connected dense units in a subspace
 - ❑ A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
 - ❑ **Subspace clustering**: A subspace cluster is a set of neighboring dense cells in an arbitrary subspace. It also discovers some minimal descriptions of the clusters
- ❑ It automatically identifies subspaces of a high dimensional data space that allow better clustering than original space using the Apriori principle

CLIQUE: The Major Steps

- Partition the data space and find the number of points that lie inside each cell of the partition.
- Identify the subspaces that contain clusters using the Apriori principle
- Identify clusters:
 - Determine dense units in all subspaces of interests
 - Determine connected dense units in all subspaces of interests.
- Generate minimal description for the clusters
 - Determine maximal regions that cover a cluster of connected dense units for each cluster
 - Determination of minimal cover for each cluster



$\tau = 3$



Strength and Weakness of *CLIQUE*

■ Strength

- It automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces
- It is *insensitive* to the order of records in input and does not presume some canonical data distribution
- It scales *linearly* with the size of input and has good scalability as the number of dimensions in the data increases

■ Weakness

- The accuracy of the clustering result may be degraded at the expense of simplicity of the method

Chapter 8. Cluster Analysis

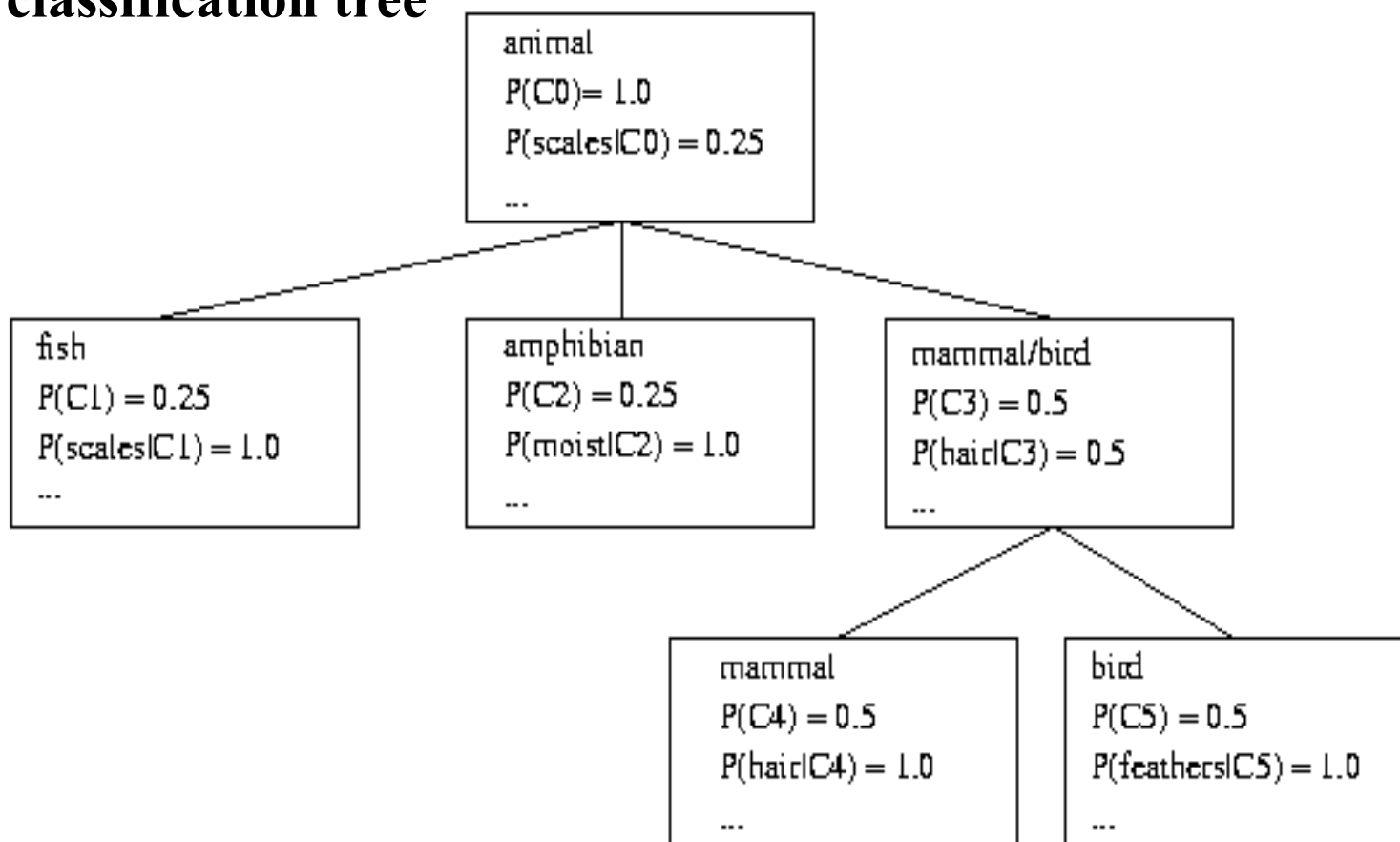
- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
 - Partitioning Methods
 - Density-Based Methods
 - Grid-Based Methods
 - Model-Based Clustering Methods
- Outlier Analysis
- Summary

Model-Based Clustering Methods

- Attempt to optimize the fit between the data and some mathematical model
- Statistical and AI approach
 - Conceptual clustering
 - A form of clustering in machine learning
 - Produces a classification scheme for a set of unlabeled objects
 - Finds characteristic description for each concept (class)
 - COBWEB (Fisher'87)
 - A popular a simple method of incremental conceptual learning
 - Creates a hierarchical clustering in the form of a **classification tree**
 - Each node refers to a concept and contains a probabilistic description of that concept

COBWEB Clustering Method

A classification tree



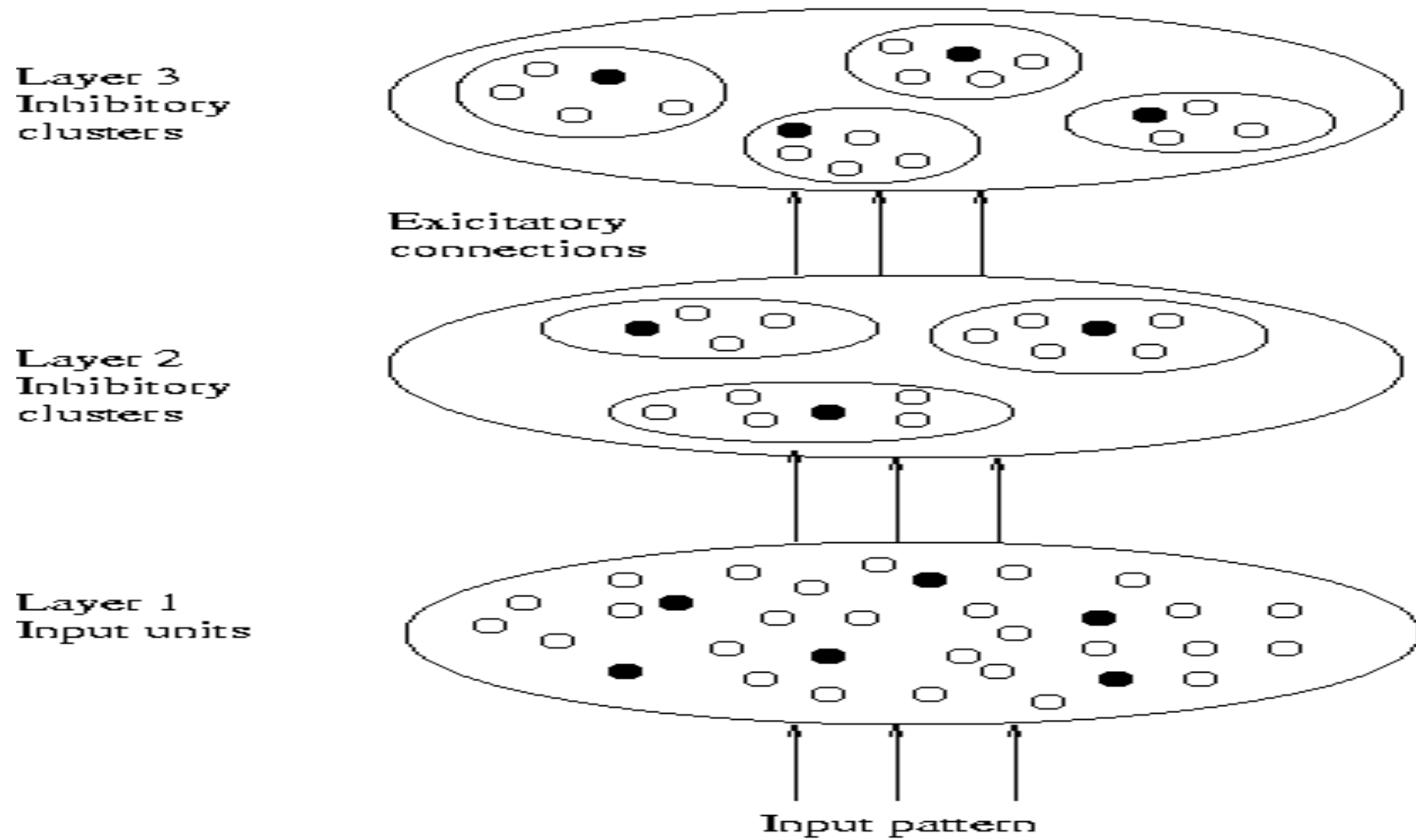
More on Statistical-Based Clustering

- Limitations of COBWEB
 - The assumption that the attributes are independent of each other is often too strong because correlation may exist
 - Not suitable for clustering large database data – skewed tree and expensive probability distributions
- CLASSIT
 - an extension of COBWEB for incremental clustering of continuous data
 - suffers similar problems as COBWEB
- AutoClass (Cheeseman and Stutz, 1996)
 - Uses Bayesian statistical analysis to estimate the number of clusters
 - Popular in industry

Other Model-Based Clustering Methods

- Neural network approaches
 - Represent each cluster as an exemplar, acting as a “prototype” of the cluster
 - New objects are distributed to the cluster whose exemplar is the most similar according to some distance measure
- Competitive learning
 - Involves a hierarchical architecture of several units (neurons)
 - Neurons compete in a “winner-takes-all” fashion for the object currently being presented

Model-Based Clustering Methods



Self-organizing feature maps (SOMs)

- Clustering is also performed by having several units competing for the current object
- The unit whose weight vector is closest to the current object wins
- The winner and its neighbors learn by having their weights adjusted
- SOMs are believed to resemble processing that can occur in the brain
- Useful for visualizing high-dimensional data in 2- or 3-D space

Chapter 8. Cluster Analysis

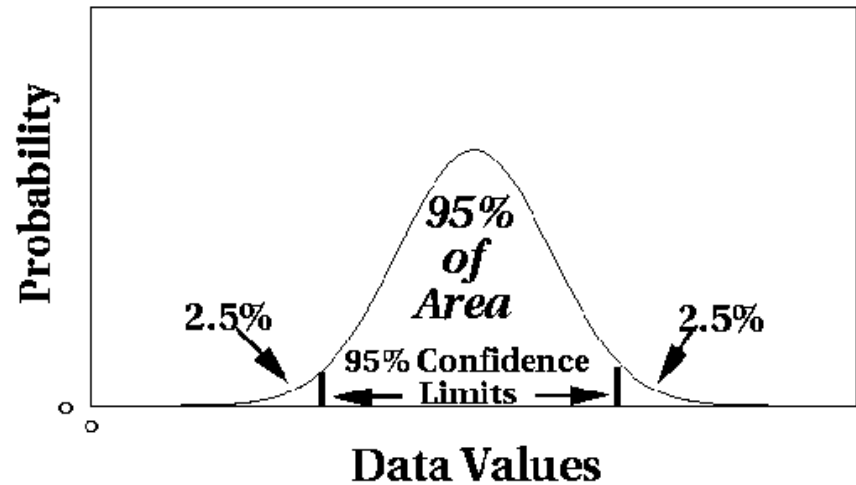
- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
 - Partitioning Methods
 - Hierarchical Methods
 - Density-Based Methods
 - Grid-Based Methods
 - Model-Based Clustering Methods
- Outlier Analysis
- Summary

What Is Outlier Discovery?



- What are outliers?
 - The set of objects are considerably dissimilar from the remainder of the data
 - Example: Sports: Michael Jordon,.....
- Problem
 - Find top n outlier points
- Applications:
 - Credit card fraud detection
 - Telecom fraud detection
 - Customer segmentation
 - Medical analysis

Outlier Discovery: Statistical Approaches



- Assume a model underlying distribution that generates data set (e.g. normal distribution)
- Use **discordancy tests to find outliers** 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

Outlier Discovery: Distance-Based Approach

- Introduced to counter the main limitations imposed by statistical methods
 - We need multi-dimensional analysis without knowing data distribution.
- Distance-based outlier: A $DB(p, D)$ -outlier is an object O in a dataset T such that at least a fraction p of the objects in T lies at a distance greater than D from O
- Algorithms for mining distance-based outliers
 - Index-based algorithm
 - Nested-loop algorithm
 - Cell-based algorithm

Outlier Discovery: Deviation-Based Approach

- Identifies outliers by examining the main characteristics of objects in a group
- Objects that “deviate” from this description are considered outliers
- sequential exception technique
 - simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects
- OLAP data cube technique
 - uses data cubes to identify regions of anomalies in large multidimensional data

Chapter 8. Cluster Analysis

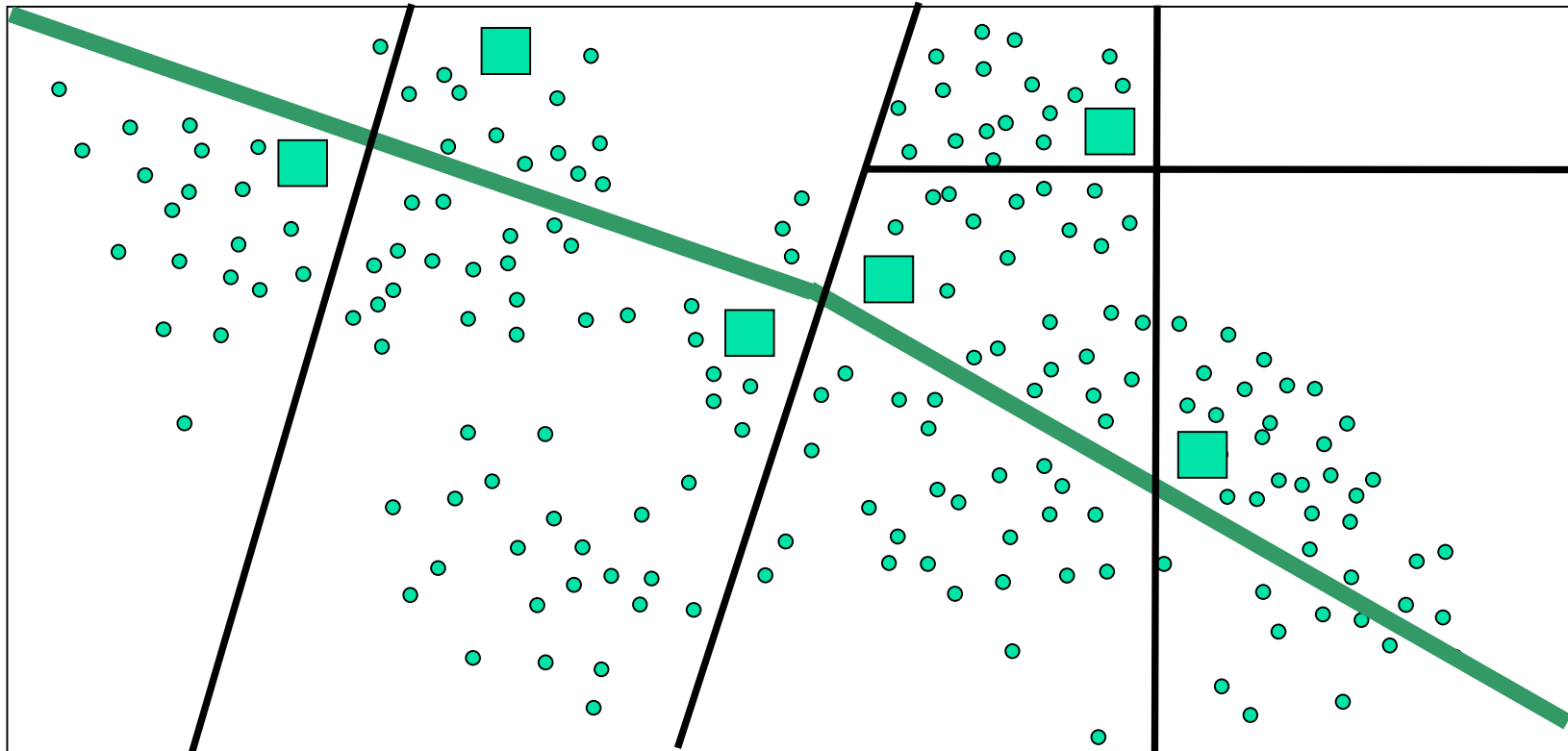
- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
 - Partitioning Methods
 - Hierarchical Methods
 - Density-Based Methods
 - Grid-Based Methods
 - Model-Based Clustering Methods
- Outlier Analysis
- Summary

Problems and Challenges

- Considerable progress has been made in scalable clustering methods
 - Partitioning: k-means, k-medoids, CLARANS
 - Hierarchical: BIRCH, CURE
 - Density-based: DBSCAN, CLIQUE, OPTICS
 - Grid-based: STING, WaveCluster
 - Model-based: Autoclass, Denclue, Cobweb
- Current clustering techniques do not address all the requirements adequately
- Constraint-based clustering analysis: Constraints exist in data space (bridges and highways) or in user queries

Constraint-Based Clustering Analysis

- Clustering analysis: less parameters but more user-desired constraints, e.g., an ATM allocation problem



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**