CLUSTER ANALYSIS

Unveiling Hidden Insights

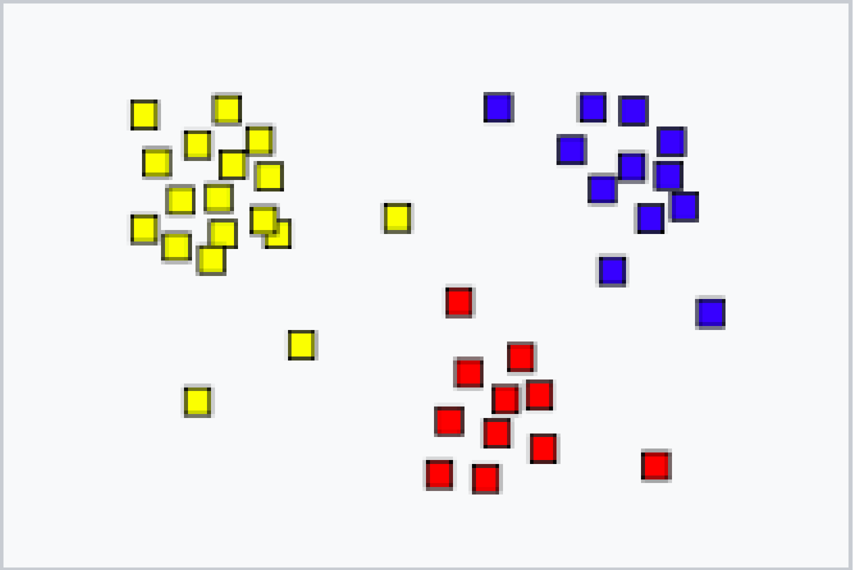
1. Harin Rishabh (50540017)
2. Tejaswini Kankanala (50539255)
3. Vaishnavi Malalur Rajegowda (50541123)

University at Buffalo The State University of New York

Contents

Introduction

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis, is a technique in data mining and statistics used to group a set of objects based on their similarities
- Objects in the same group are more similar to each other than to those in other groups.
- It's a method of unsupervised learning, which means it doesn't rely on pre-labeled data. Instead, it identifies patterns and structures within the data on its own.
- Figure: It depicts how the similar objects are grouped into different clusters.



The result of a cluster analysis shown as the coloring of the squares into three clusters

Purpose of Cluster Analysis

- "The primary goal of clustering is to identify patterns or structures within datasets without prior knowledge of group memberships, facilitating exploratory data analysis, summary generation, and outlier detection."
- **Primary Goals of Cluster Analysis**

  i. Identifying Patterns or Structures

  ii. Facilitating Exploratory Data Analysis

  iii. Summary Generation

  iv. Outlier Detection

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 marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earthquake epicenters should be clustered along continent faults

What is a 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

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.

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

Data matrix
(two modes)

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

Dissimilarity matrix
(one mode)

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

- 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

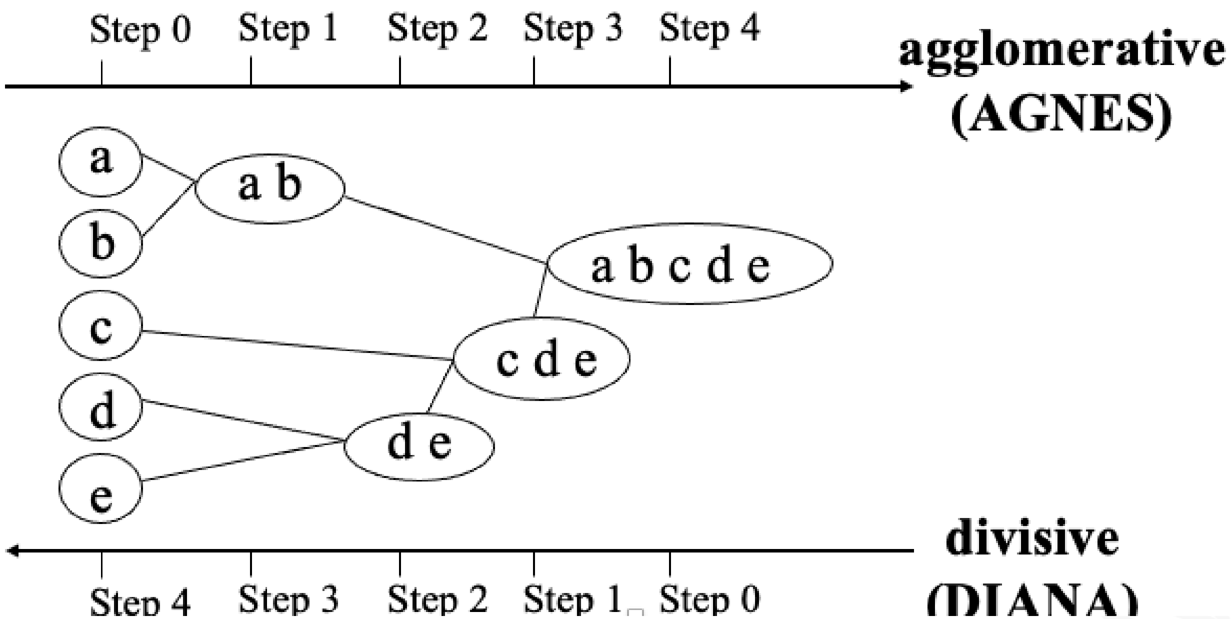
- Given *k*, the *k-means* algorithm is implemented in 4 steps:
- Partition objects into *k* nonempty subsets
- Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
- Assign each object to the cluster with the nearest seed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

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

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

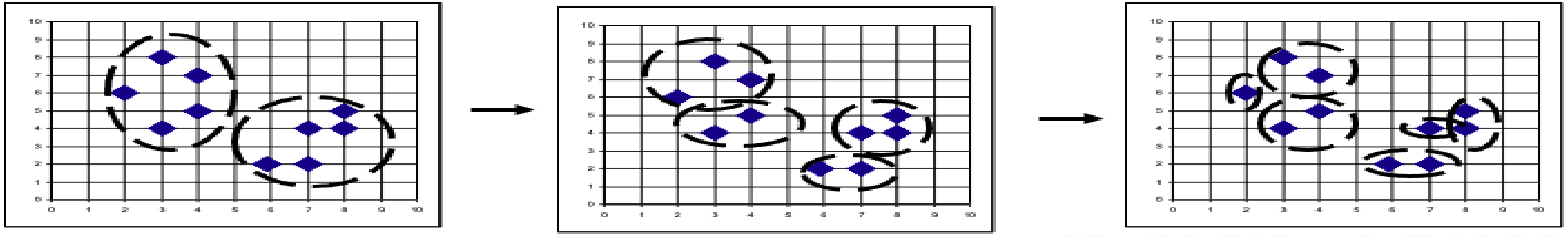


AGNES (Agglomerative Nesting)

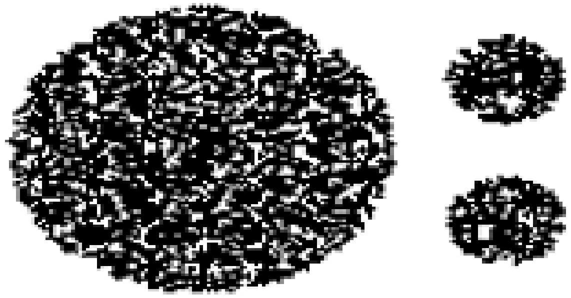
- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Use the Single-Link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster

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



CURE (Clustering Using REpresentatives)



(a)

(b)

CURE: proposed by Guha, Rastogi & Shim, 1998

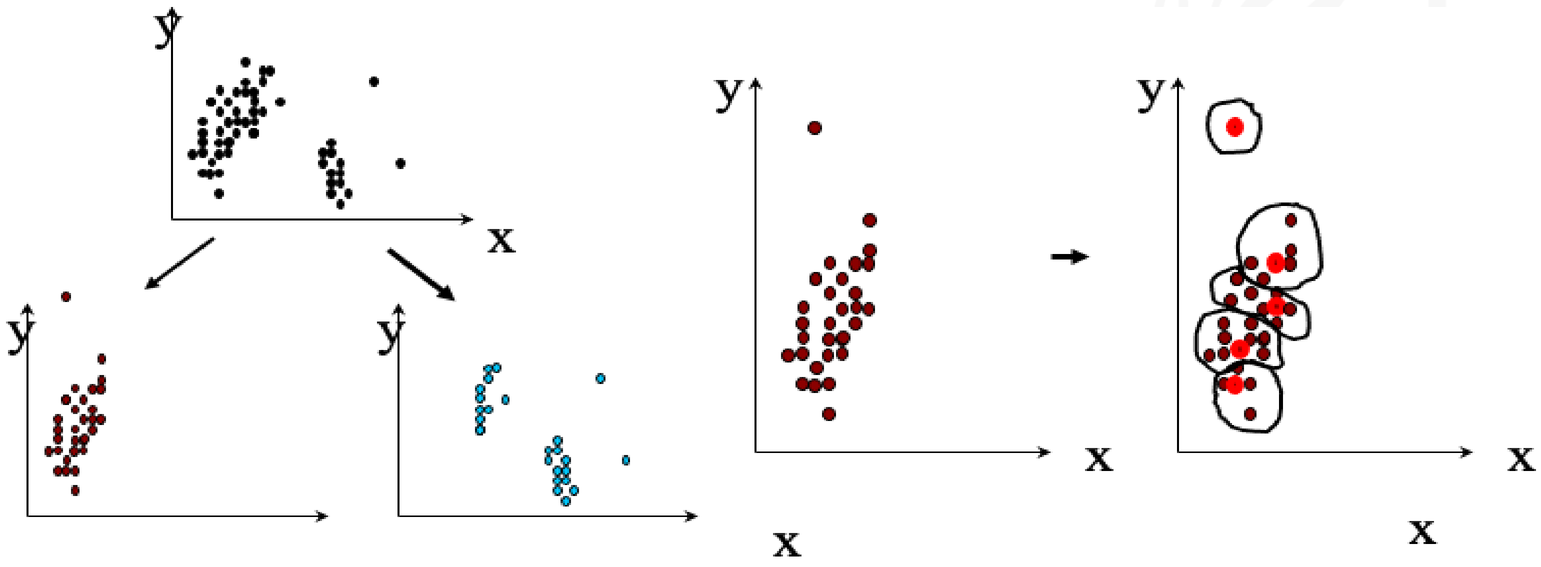
- Stops the creation of a cluster hierarchy if a level consists of *k* clusters
- Uses multiple representative points to evaluate the distance between clusters, adjusts well to arbitrary shaped clusters and avoids single-link effect

CURE: The Algorithm

- Draw random sample *s*.
- Partition sample to *p* partitions with size *s/p*
- Partially cluster partitions into *s/pq* clusters
- Eliminate outliers
- By random sampling
- If a cluster grows too slow, eliminate it.
- Cluster partial clusters.
- Label data in disk

Data Partitioning and Clustering

- ns = 50
- np = 2. ns/pq = 5
- ns/p = 25



Rock: Algorithm

- Links: The number of common neighbors for the two points.

{1,2,3}, {1,2,4}, {1,2,5}, {1,3,4}, {1,3,5}

{1,4,5}, {2,3,4}, {2,3,5}, {2,4,5}, {3,4,5}.

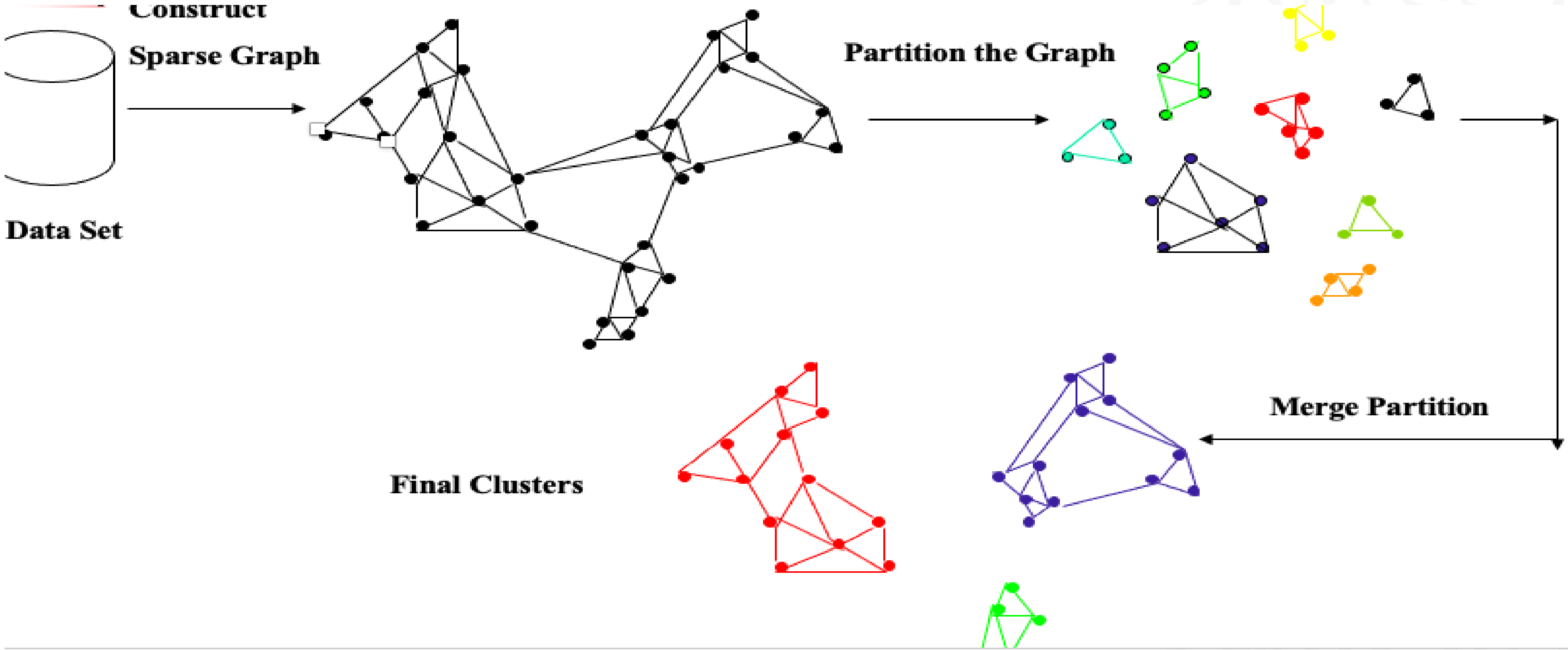
Algorithm

- Draw random sample
- Cluster with links
- Label data in disk

CHAMELEON

- CHAMELEON: hierarchical clustering using dynamic modeling, by G. Karypis, E.H. Han and V. Kumar '99
- Measures the similarity based on a dynamic model
- Two clusters are merged only if the *interconnectivity* and *closeness (proximity)* between two clusters are high *relative to* the internal interconnectivity of the clusters and closeness of items within the clusters
- A two phase algorithm
- 1. Use a graph partitioning algorithm: cluster objects into a large number of relatively small sub-clusters

Overall Framework of CHAMELEON



Density-Based Clustering Methods

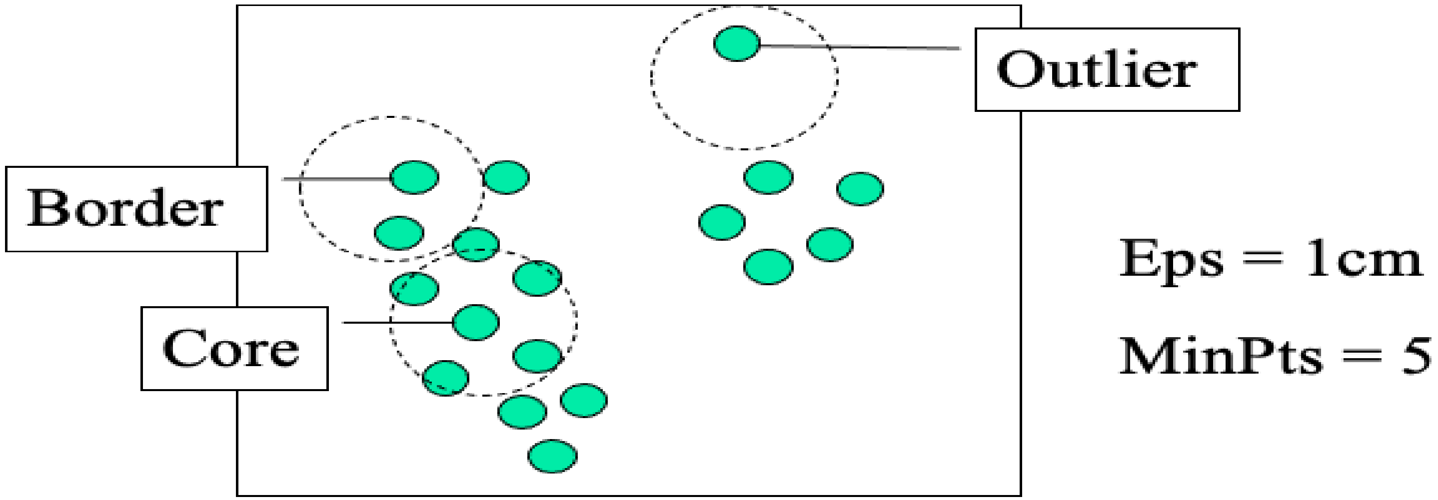
- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
- Discover clusters of arbitrary shape
- Handle noise
- One scan
- Need density parameters as termination condition

Density-Based Clustering Methods

- 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)

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

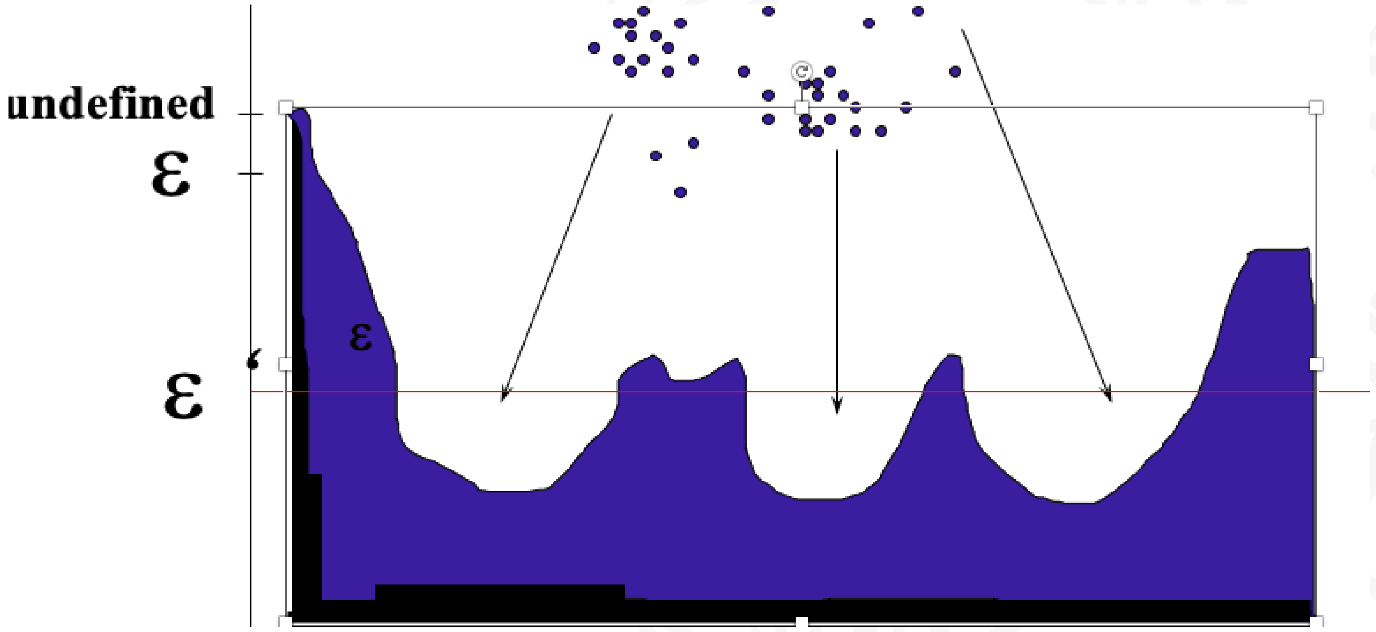


DBSCAN: The Algorithm

- 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 atabase.
- Continue the process until all of the points have been processed.

OPTICS: Some Extension from DBSCAN

- Index-based:
- k = number of dimensions
- N = 20
- p = 75%
- M = N(1-p) = 5
- Complexity: $O(kN^2)$

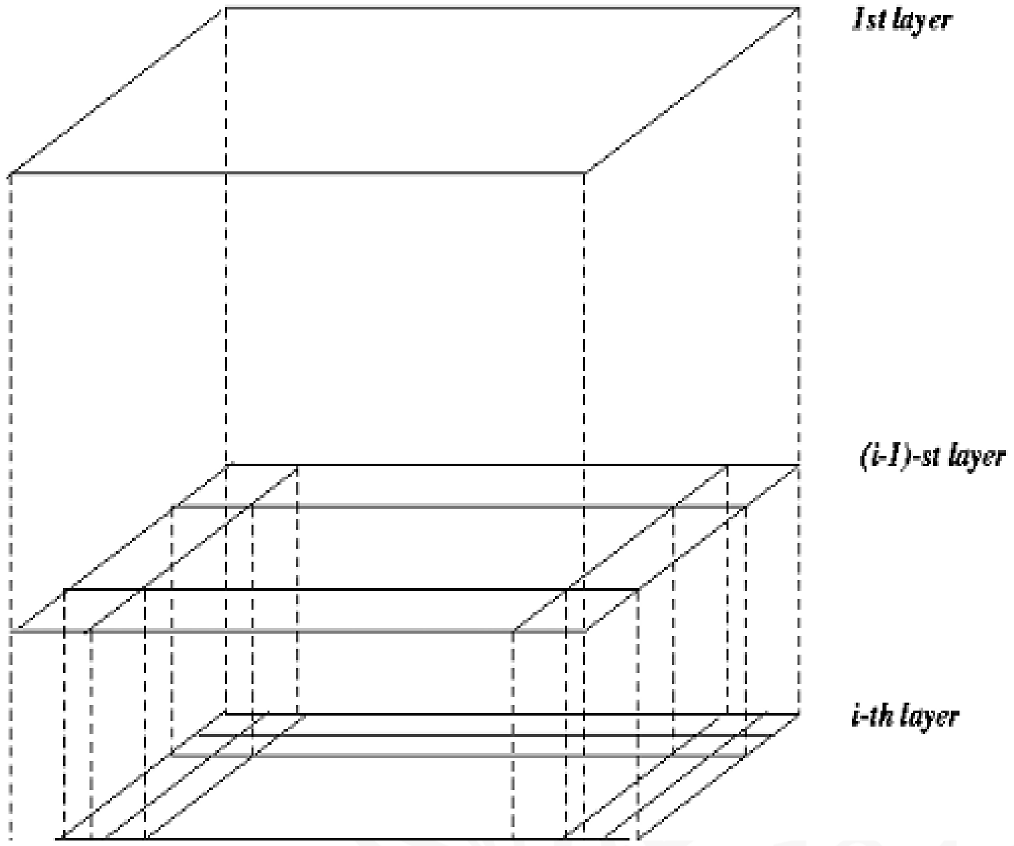


Grid-Based Clustering Method

- 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

STING: A Statistical Information Grid Approach

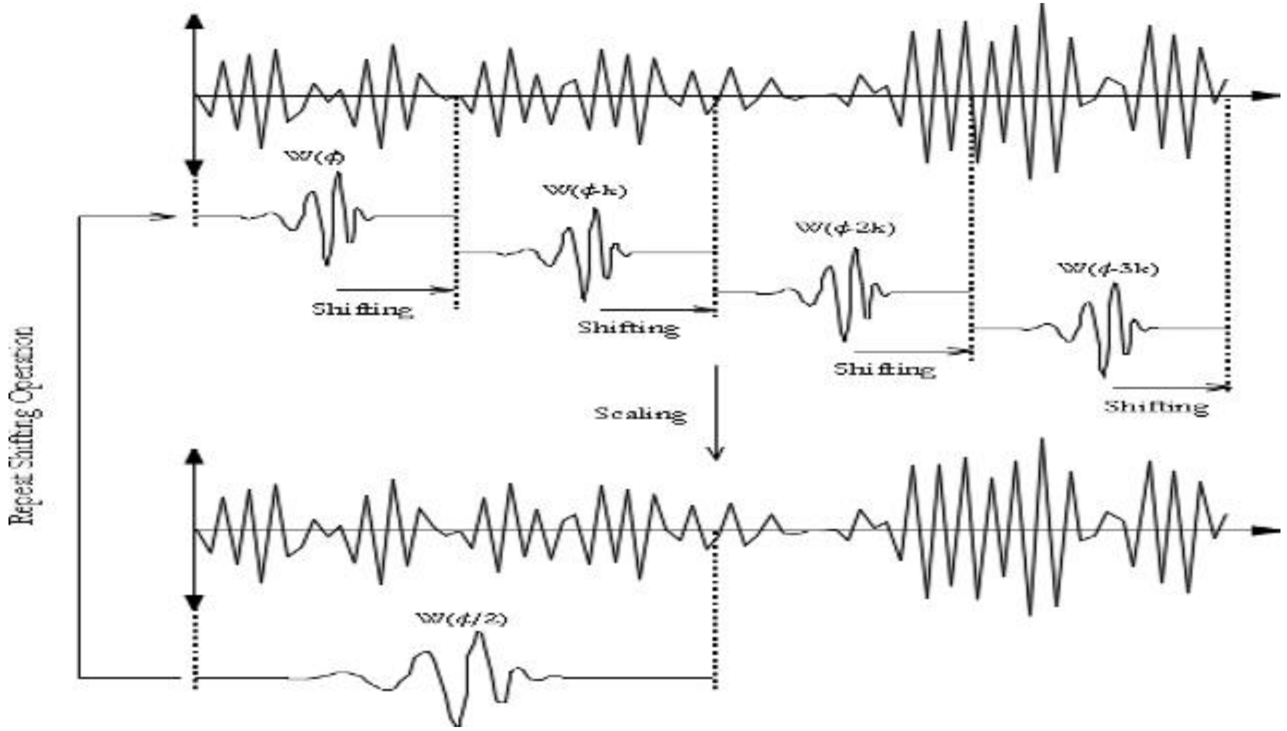
- Wang, Yang and Muntz (VLDB’97)
- The spatial area area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution



WaveCluster

- Sheikholeslami, Chatterjee, and Zhang (VLDB’98)
- 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

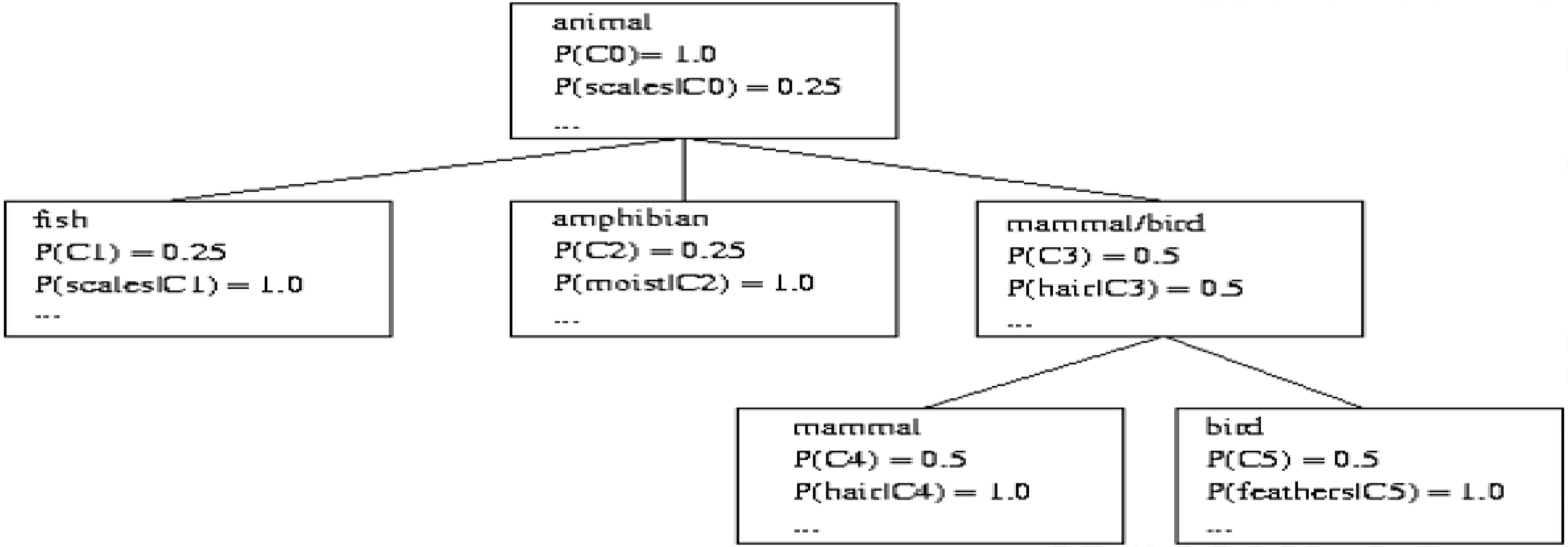
Wavelet cluster



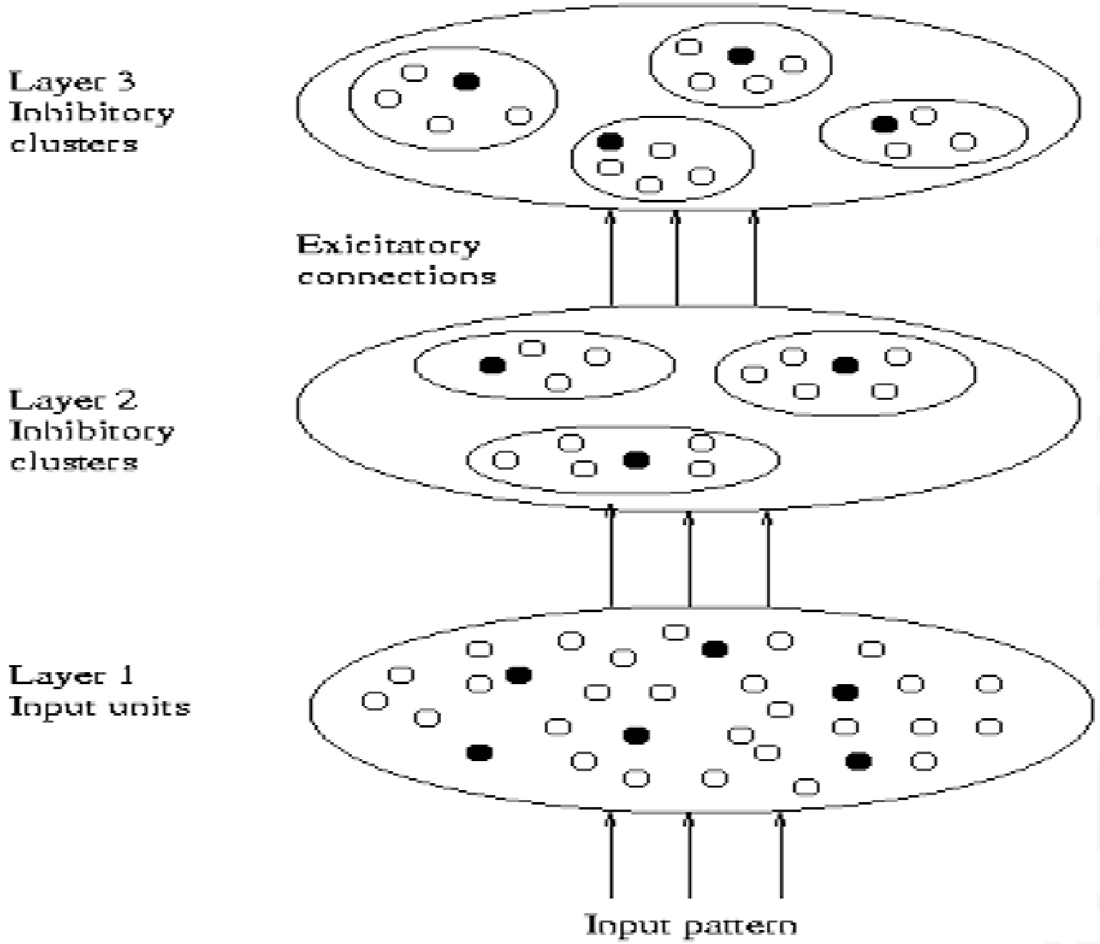
Model-Based Clustering Methods

- 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

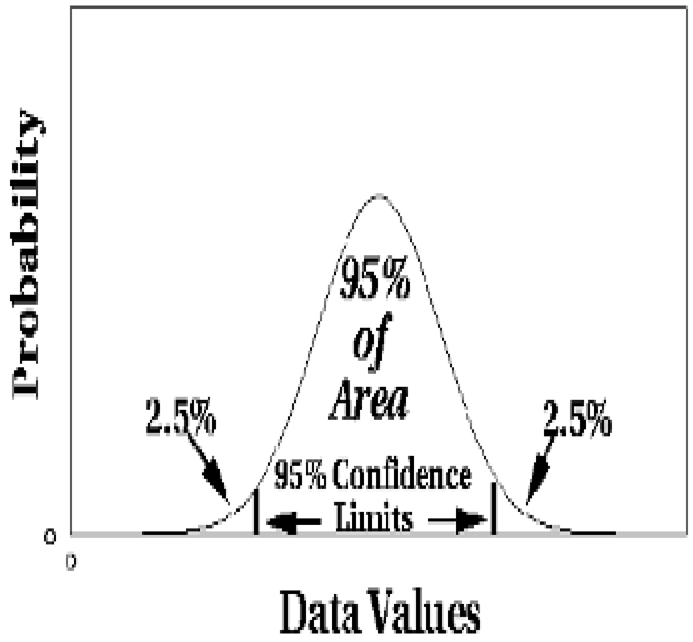


Model-Based Clustering Methods



Outlier Discovery: Statistical Approaches

- 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



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

References

- RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS by Amit Mishra link
- A detailed study of clustering algorithms by IEEE link
- Cluster analysis: A modern statistical review by Adam Jaeger link
- K-Means Cluster Analysis link
- Review Paper on Clustering Techniques by Amandeep Kaur Mann & Navneet Kaur link

THANK YOU