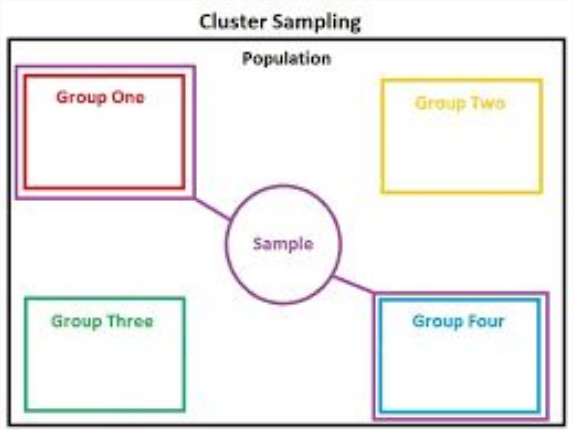
CLUSTER ANALYSIS

Agenda

- Introduction to cluster analysis and application
- Feature Selection for Clustering
- Clustering Algorithm
- DBScan
- Cluster Validation

Introduction

Clustering is the partitioning of large number of data into smaller number of groups.



Applications

- Data Summarization
- Customer Segmentation
- Social Network Analysis
- Relationship to other data mining problems

Feature Selection in Clustering

The aim of feature selection is to remove noisy attributes that do not cluster well. Difficult in clustering because it is unsupervised.

- Filter Model
- Wrapper Model

Unique to each student

Student ID	Name	Score	Class
T00525452	John	80	Grade 5
T00423514	Peter	79	Grade 5
T00321452	James	55	Grade 3
T00715261	Tony	63	Grade 4

Filter Model and Wrapper Model

In filter model, a specific criterion is used filter out features or data points that does not meet specific score and affect clustering tendency.

In wrapper model, clustering algorithm is used to evaluate quality of features

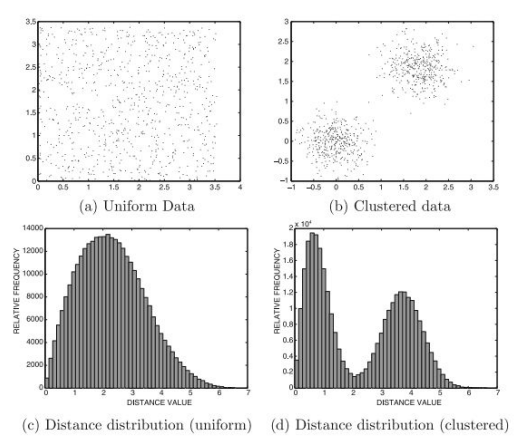
Wrapper Model - Integrated into clustering process

Filter Model - preprocessing phase

Common Filter Model Criterions

- **Term Strength** - fraction of similar document pair (sparse domain)
- **Predictive Attribute Dependence** - correlated features form better clusters
  - Classification Algorithm
  - Regression Modeling (Numeric)
- **Entropy -** clustered data reflect clustering characteristics.
- **Hopkins statistic-** measure clustering tendency of

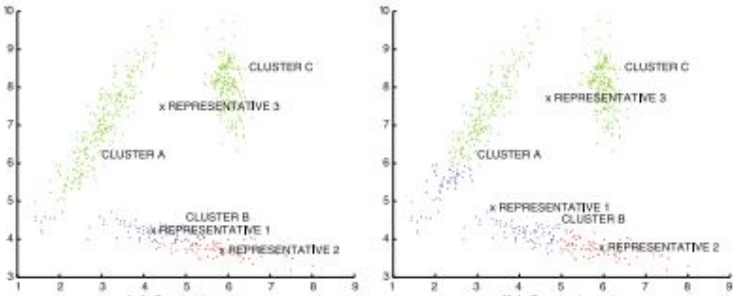
  Data set



CLUSTERING ALGORITHM

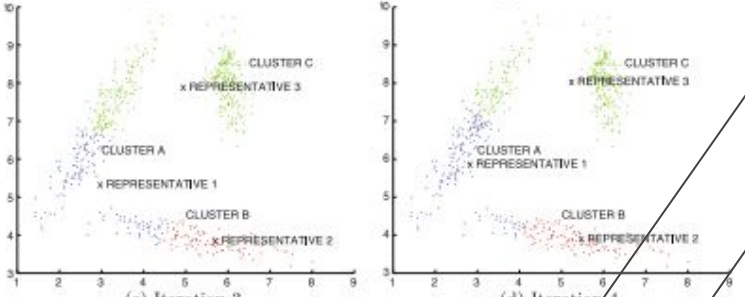
Representative-Based Algorithm

- Simplest clustering algorithm
- No hierarchical relationship
- Cluster created in one shot
- Uses partition representative



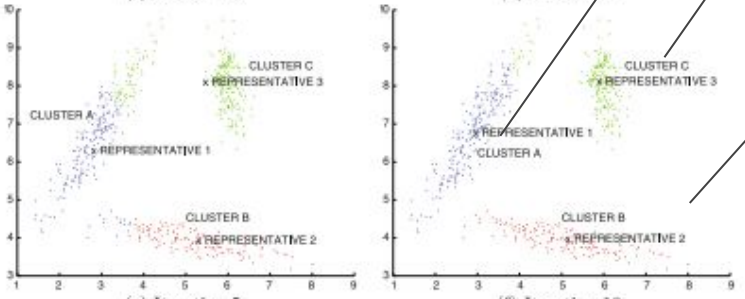
(a) Iteration 1

(b) Iteration 2



(c) Iteration 3

(d) Iteration 4



(e) Iteration 5

(f) Iteration 10

Figure 6.3: Illustration of k-representative algorithm with random initialization

Minimization of distance of the different data point and the closest representative

Types of Representative Based Algorithm

- k-Means Algorithm - Euclidean Distance, does not work when cluster are arbitrary of shape
- Kernel K-Means - determine arbitrary shape of cluster
- k-Median - Manhattan distance
- K-medoids - representatives selected from database because k-means can be distorted by outliers

Hierarchical Clustering Algorithm

- Cluster data with distance
- Distance function not compulsory
- Density based or graph based algorithm can be used

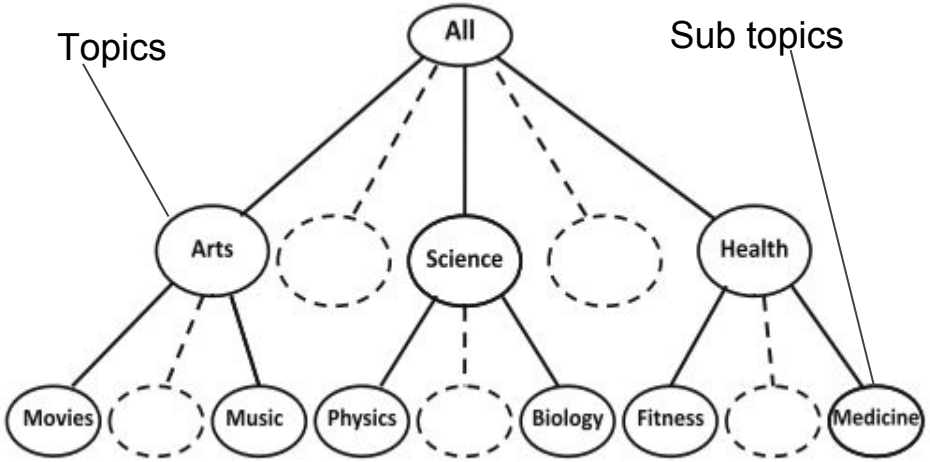
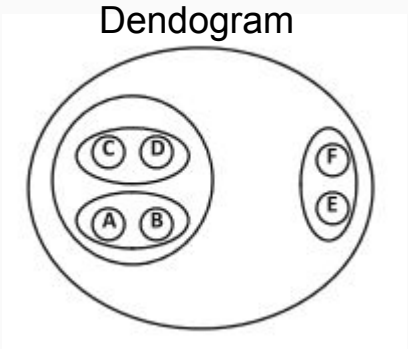
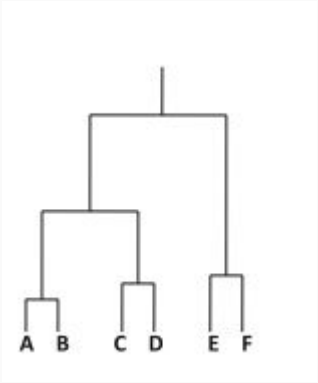


Figure 6.6: Multigranularity insights from hierarchical clustering

Types of Hierarchical Clustering Algorithm

- Bottom-up (agglomerative) Methods - individual data points are agglomerated to higher level
- Top Down (divisive) Methods - partition data point into three like structures





Probabilistics Model-Based Algorithm

- Soft algorithm
- Each data point may have a nonzero assignment probability to many clusters
- Soft solution can be converted to hard solution by assignment of data point to each cluster

Grid-Based and Density-Based Algorithm

Grid Based

- Number of data cluster is not pre-defined in advanced compared to k-means
- Return natural cluster with in data with corresponding shapes

Density Based Clustering Method

- Clustering based on density such as density connected points or based on an explicitly constructed density function
- Can handle noise
- Handles clusters of all shapes
- One scan
- Doesn’t work well varying densities and high dimensional data

DBScan

- DBScan is a density based algorithm, used to find associations and structures in data that are hard to find manually but can be relevant to find patterns and predict trends
- Density = number of points within a specified radius(Eps)
- A point is a core point if it has more than a specified number of points(MinPts) within Eps
- A border point has fewer MinPts within Eps, but is in the neighbourhood of a core point
- A noise point is any point that is not a core point or a border point

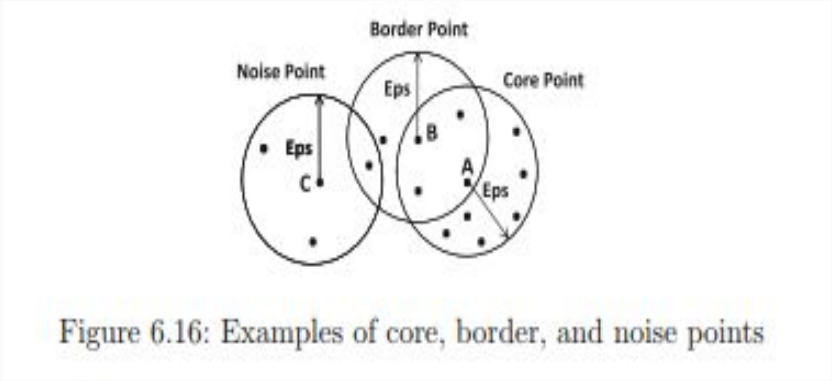
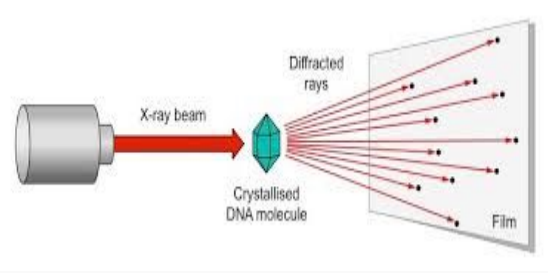


Figure 6.16: Examples of core, border, and noise points

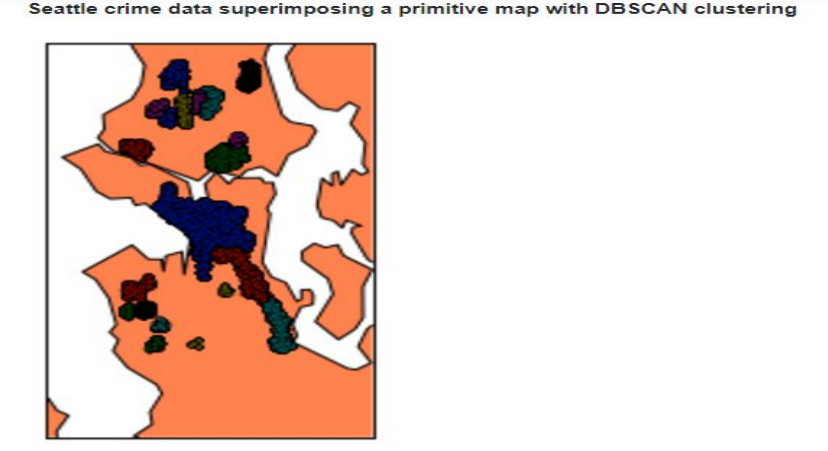
Practical Applications of DBScan

- Satellite Imagery
- X-ray Crystallography
- Anomaly Detection in Temperature Data
- Credit Fraud Detection
- E-commerce (recommending relevant products to customers)

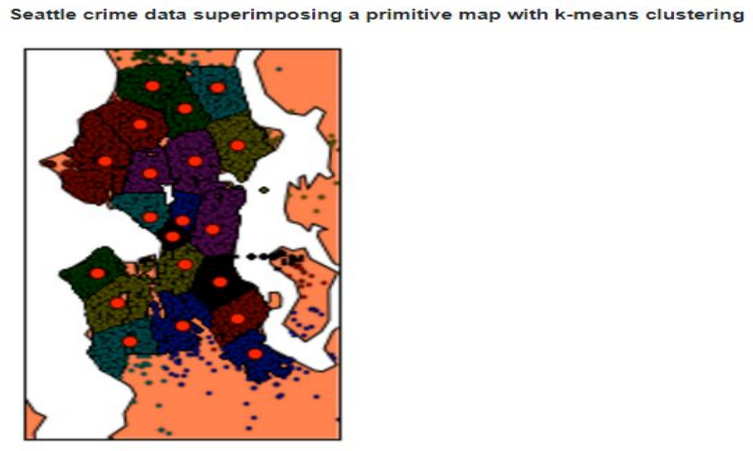


DBScan vs K-means clustering

- In terms of location analytics, Dbscan does a wonderful job of identifying high crime areas through density filtering,

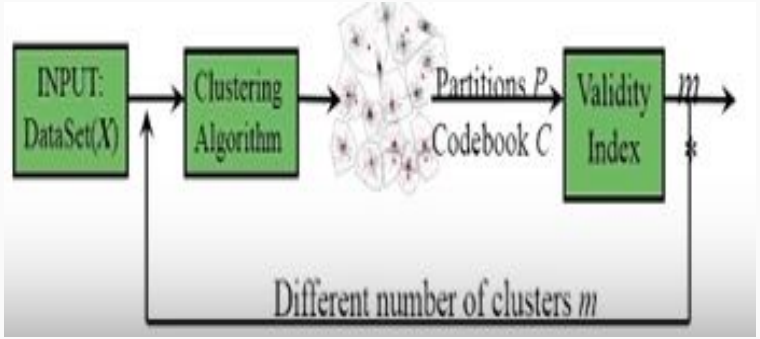


- K-means clustering works fine but doesn't provide additional insight



Cluster Validation

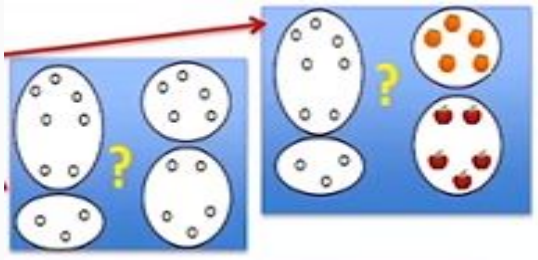
- Cluster validation refers to procedures that evaluate the results of clustering in a quantitative and objective fashion. [Jane & Dubes, 1988]
- Quantitative means to employ the measures while objective means to validate the measures
- Cluster validation is needed to compare cluster algorithms, solve number of clusters and avoid finding patterns in Noise



Measures of Cluster Validation

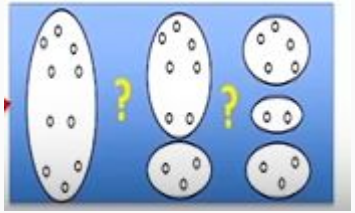
External Index

- Comparing the clustering results to ground truth (externally known results)
- Comparing two clusters



Internal Index

- Evaluating quality of clusters without reference to external information
- Validate with different number of clusters



Thank you