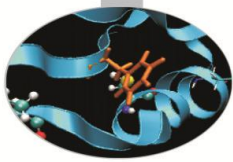
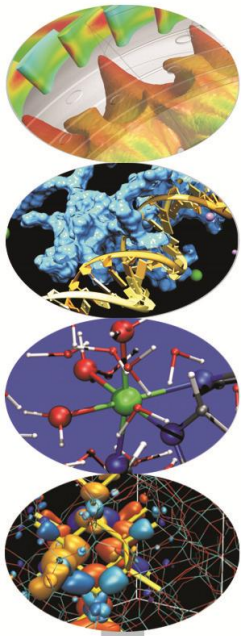
CINECA SCAI
SuperComputing Applications and Innovation



Unsupervised techniques



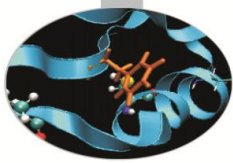
Giorgio Pedrazzi, *CINECA-SCAI*

School of Data Analytics and Visualisation

Milan, 09/06/2015

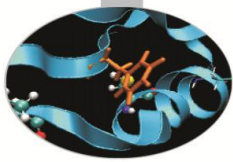
CINECA

Unsupervised Learning



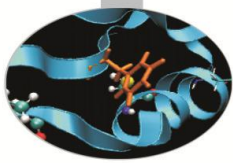
- Cluster Analysis
  - Basic concept and introduction
  - Major clustering approaches
  - Distance measures
- Association rules and sequential patterns
  - Association rules
  - Sequential patterns

Introduction



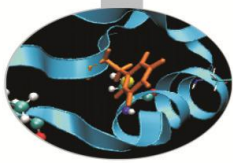
- Cluster: A collection/group of data objects/points
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis
  - find *similarities* between data according to characteristics underlying the data and grouping similar data objects into clusters
- Clustering Analysis: Unsupervised learning
  - no predefined classes for a training data set
  - Two general tasks: identify the “natural” clustering number and properly grouping objects into “sensible” clusters
- Typical applications
  - as a stand-alone tool to gain an insight into data distribution
  - as a preprocessing step of other algorithms in intelligent systems

Introduction



- The problem must be formulated in a mathematical way as a matrix of data containing information on N objects (cases or observations ; rows of the matrix) specified by the values assigned to V variables (columns of the matrix)

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

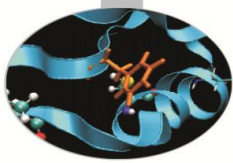


Cluster Analysis steps

- Pre processing
- Select a clustering algorithm
- Select a distance or a similarity measure (*)
- Determine the number of clusters (*)
- Validate the analysis

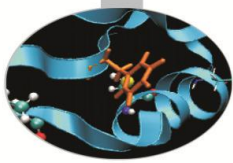
(*) if needed by the method used

Classification of methods

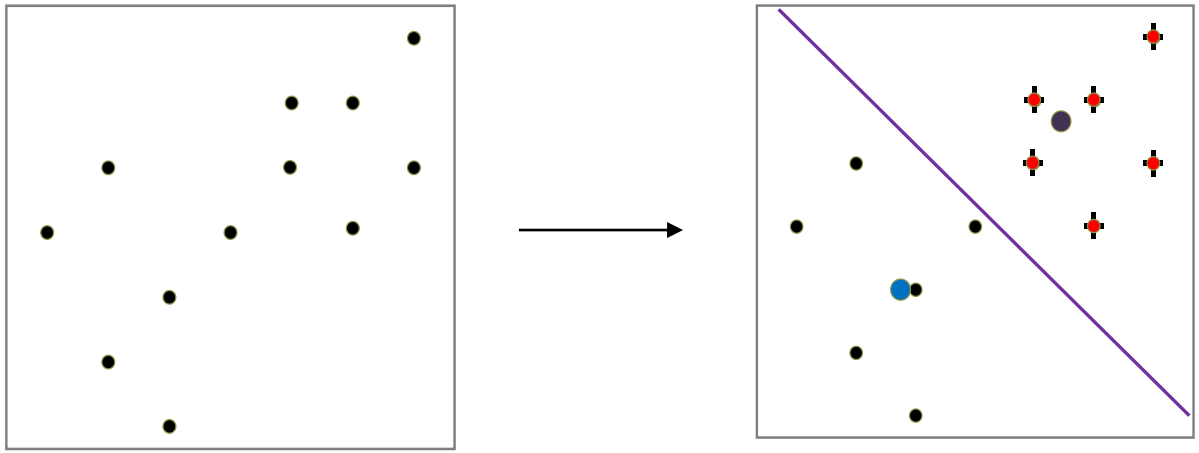


- Distance-based
  - partitioning algorithms
  - hierarchical algorithms
- Density based
- Model based
- Spectral
- Combination of methods

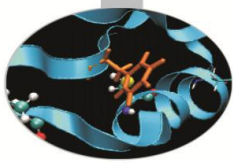
Partioning Approach



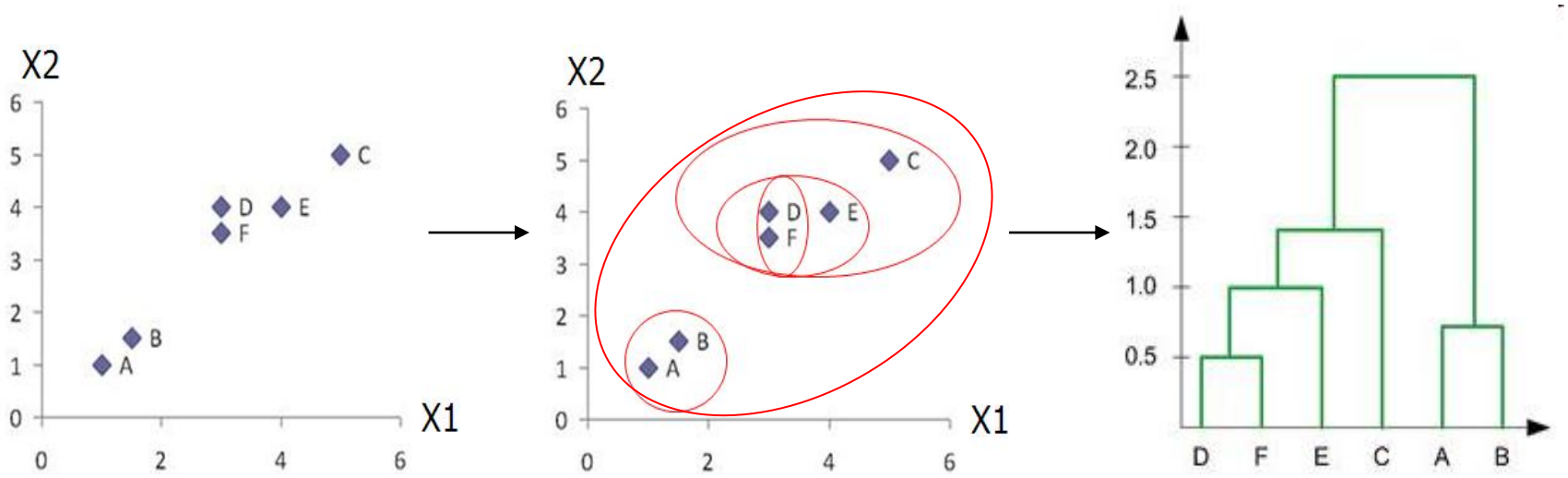
- Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square distance cost
- Typical methods: K-Means, K-Medoids, K-Medians, ……



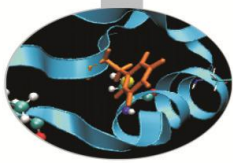
Hierarchical Approach



- Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Typical methods: Agglomerative, Diana, Agnes, BIRCH, ROCK, ……



Distance measure



Minkowski distance (L_p Norm)

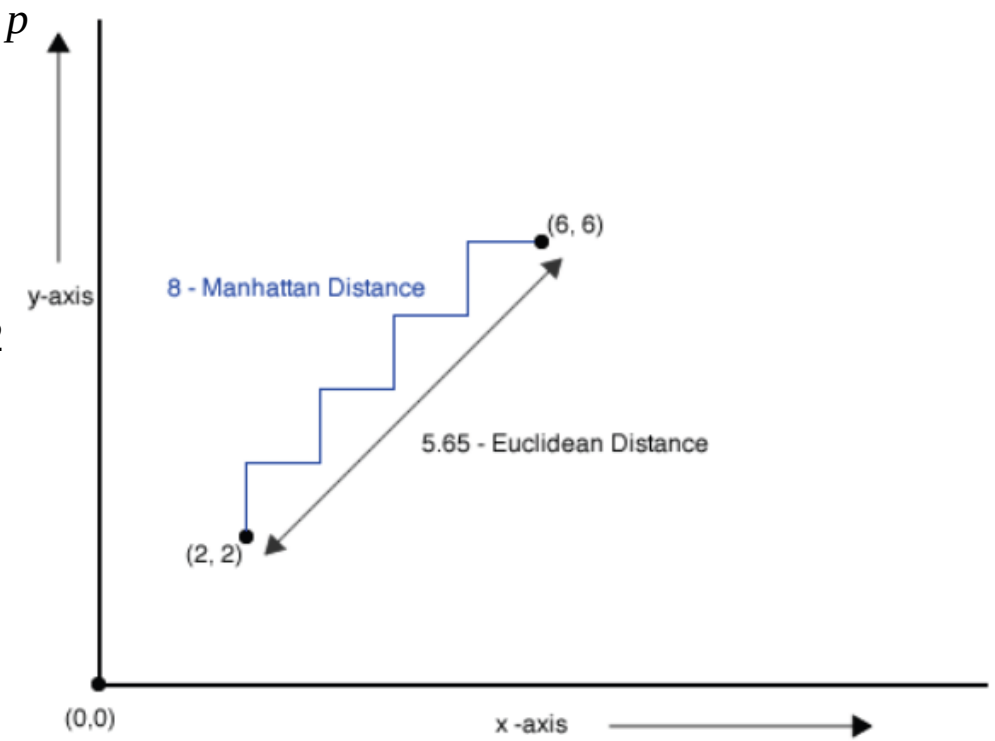
$$d(i,k) = \left[\sum_{j=1}^{d} | x_{ij} - x_{kj} |^{p} \right]^{1/p}$$

Euclidean distance (L_2 Norm)

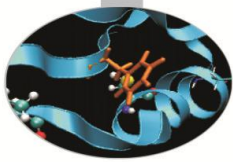
$$d(i,k) = \left[\sum_{j=1}^{d} (x_{ij} - x_{kj})^{2} \right]^{1/2}$$

Manhattan distance (city block distance)

$$d(i,k) = \sum_{j=1}^{d} | x_{ij} - x_{kj} |$$



Distance Measures



- Cosine Measure (Similarity vs. Distance)

For $\mathbf{x} = (x_1\, x_2 \cdots x_n)$ and $\mathbf{y} = (y_1\, y_2 \cdots y_n)$

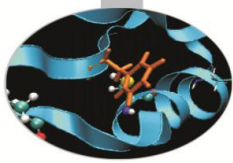
$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \bullet \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{x_1 y_1 + \cdots + x_n y_n}{\sqrt{x_1^2 + \cdots + x_n^2}\sqrt{y_1^2 + \cdots + y_n^2}}$$

$$d(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \bullet \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

$$0 \le d(\mathbf{x}, \mathbf{y}) \le 2$$

  - Nonmetric vector objects: keywords in documents, gene features in micro-arrays, …
  - Applications: information retrieval, biologic taxonomy, ...

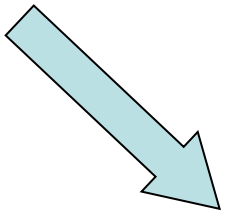
Similarity measures

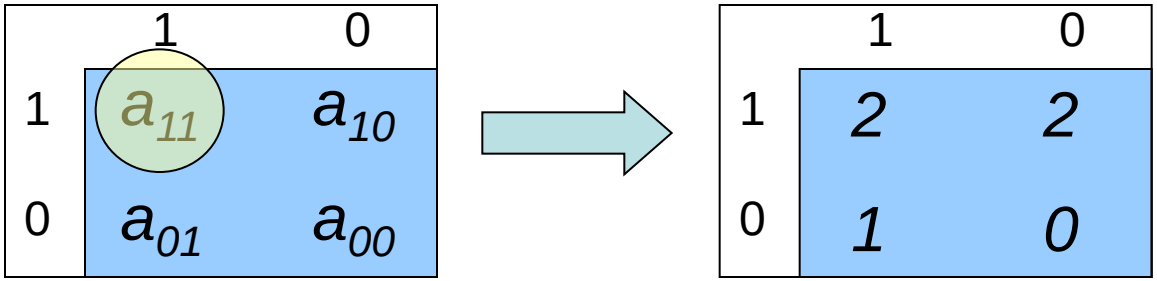


Correspondent 1's

x_k: 0 1 1 0 1

x_j: 1 1 0 1 1



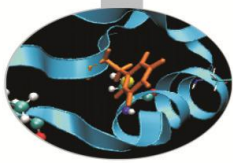


Jaccard: $d(i,k) = (a_{11}) / (a_{11} + a_{10} + a_{01})$

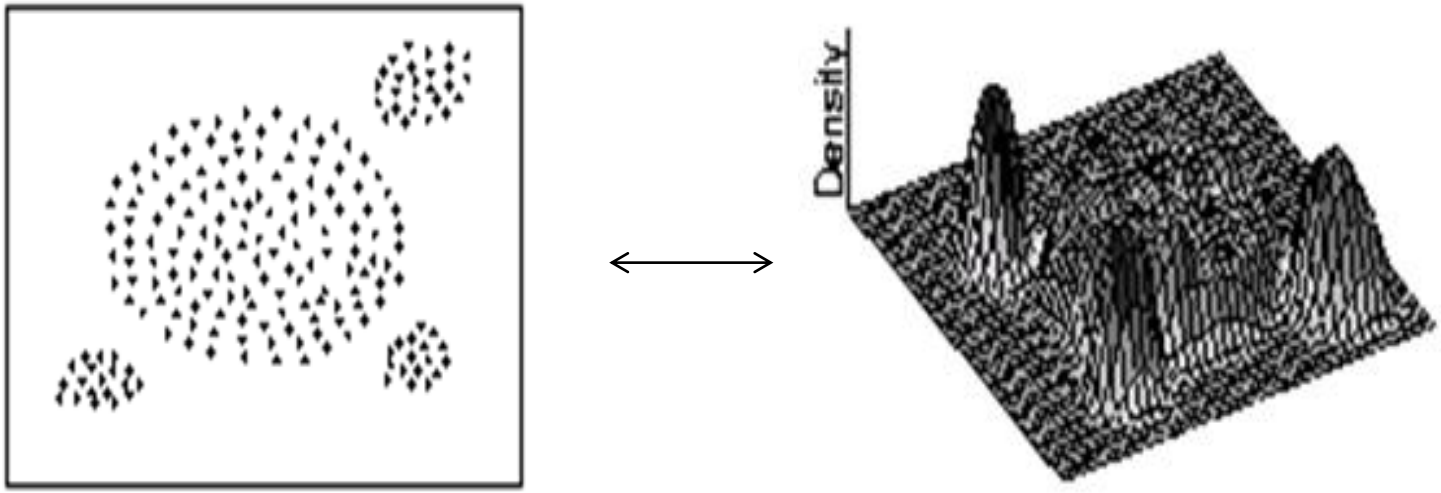
Condorcet: $d(i,k) = a_{11} / [a_{11} + 0.5(a_{10} + a_{01})]$

Dice bis: $d(i,k) = a_{11} / [a_{11} + 0.25(a_{10} + a_{01})]$

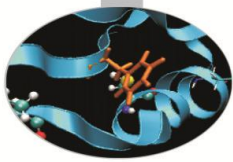
Density based Approach



- Based on connectivity and density functions
- Typical methods: DBSCAN, OPTICS, DenClue, ……

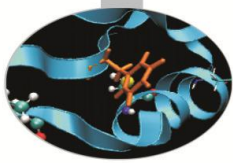


Density based approach

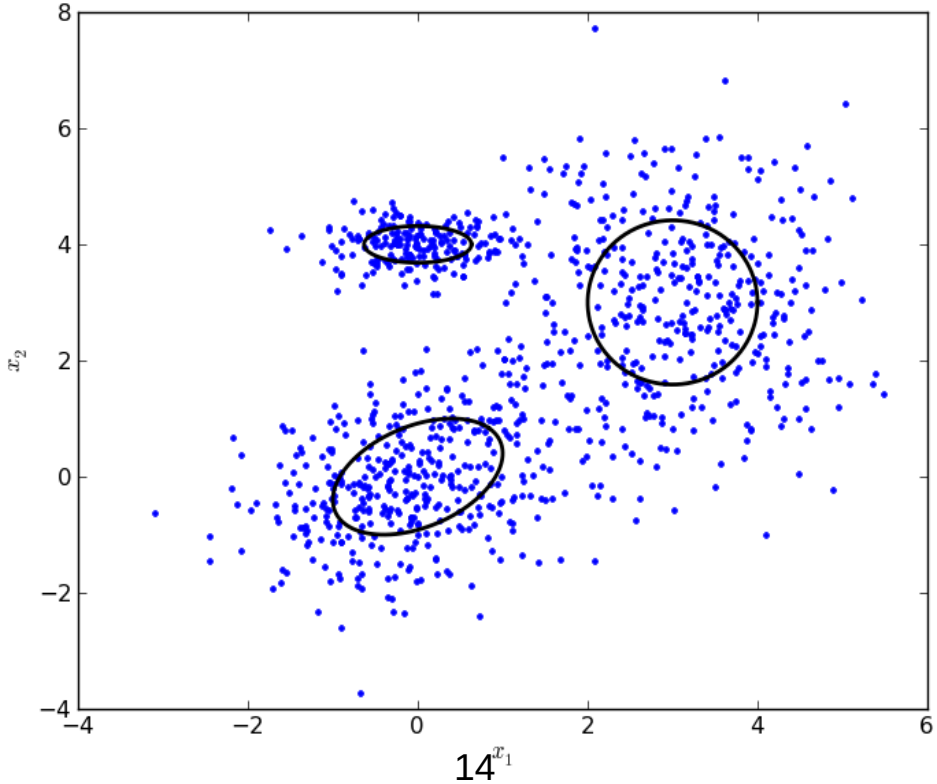


Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature

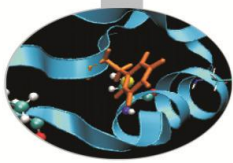
Model-based Approach



- For each cluster, a theoretical model is hypothesized in order to find the best fit.
- Typical methods: Gaussian Mixture Model (GMM), COBWEB, ……

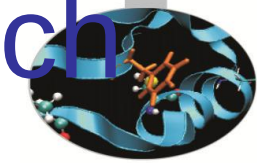


Model-based Approach

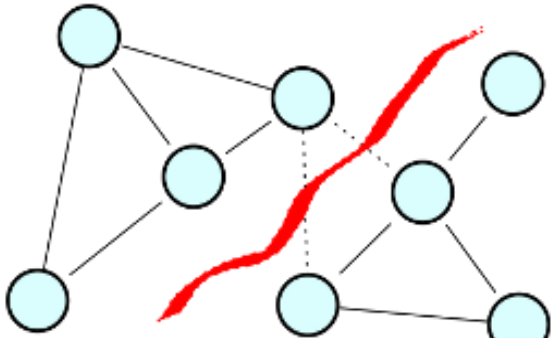


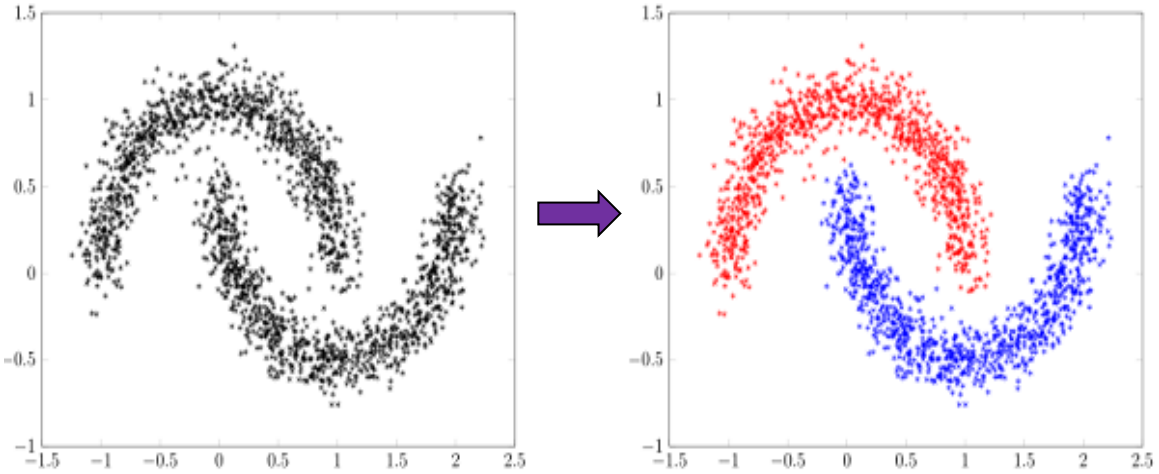
- Probabilistic model-based clustering
  - In model-based clustering, it is assumed that the data are generated by a mixture of underlying probability distributions in which each component represents a different group or cluster.
  - Cluster: Data points (or objects) that most likely belong to the same distribution
  - Clusters are created so that they will have a maximum likelihood fit to the model by a mixture of K component distributions (i.e., K clusters)

Spectral Clustering Approach

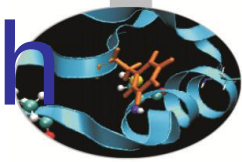


- Convert data set into weighted graph (vertex, edge), then cut the graph into sub-graphs corresponding to clusters via spectral analysis
- Typical methods: Normalised-Cuts ……





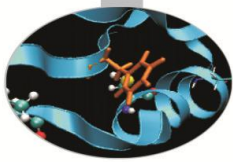
Spectral Clustering Approach



- In multivariate statistics, spectral clustering techniques make use of eigenvalue decomposition (spectrum) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset.

- In application to image segmentation, spectral clustering is known as segmentation-based object categorization.

Combination of methods

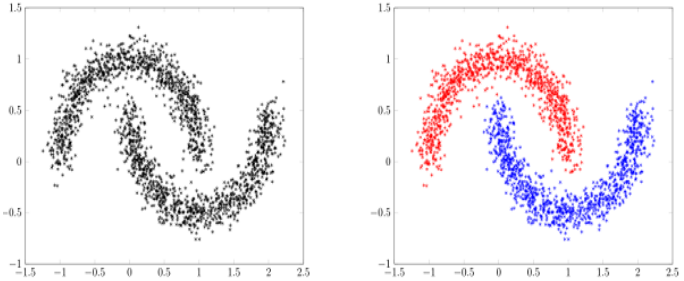


Using different methods can be useful for overcome the drawbacks of a single methods.

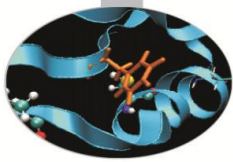
For example it is possible to generate a large number of clusers with K-means and then cluster them together using a hierarchical method.

It is important using the “single-link” method, in which the distance between two clusters is defined by the distance between the two closest data points we can find, one from each cluster.

This method has been applied to find cluster in non-convex set.



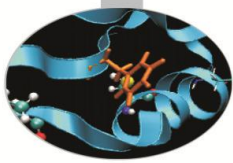
Clustering validation



Clustering validation is a technique to find a set of clusters that best fits natural partitions (number of clusters) without any class information. Generally speaking, there are two types of clustering techniques, which are based on external criteria and internal criteria.

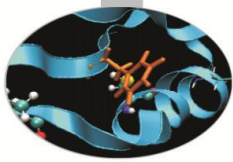
- External validation: Based on previous knowledge about data (F-measure, NMIMeasure, Entropy, Purity)
- Internal validation: Based on the information intrinsic to the data alone (BIC, CH, DB, SIL, DUNN)

Summary



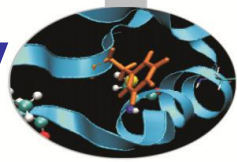
- Clustering analysis groups objects based on their (dis)similarity and has a broad range of applications.
- Measure of distance (or similarity) plays a critical role in clustering analysis and distance-based learning.
- Clustering algorithms can be categorized into partitioning, hierarchical, density-based, model-based, spectral clustering as well as combination approaches.

Market Basket Analysis



- Basket data consist of collection of transaction date and items bought in a transaction
- Retail organizations interested in generating qualified decisions and strategy based on analysis of transaction data
  - what to put on sale, how to place merchandise on shelves for maximizing profit, customer segmentation based on buying pattern
- Examples.
  - Rule form: LHS $\rightarrow$ RHS [confidence, support].
  - diapers $\rightarrow$ beers [60%, 0.5%]
  - “90% of transactions that purchase bread and butter also purchase milk”
  - bread and butter $\Rightarrow$ milk [90%, 1%]

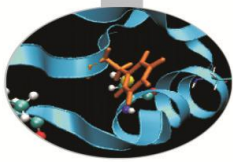
Association Rules discovery



- The problem of mining association rules:

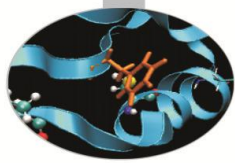
  Generate all association rules that have support and confidence greater than the user-specified *minsup* and *minconf*
- **Minimum support s** (*minsup*) --- the union of items in the LHS and RHS of the rule is present in a minimum of s% of transactions in the database
- **Minimum confidence c** (*minconf*) --- at least c% of transactions in the database that satisfy the LHS of the rule also satisfy the RHS of the rule

Association rule discovery problem



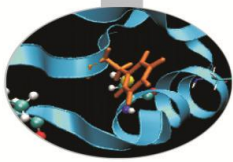
- Two sub-problems in discovering all association rules:
  - Find all sets of items (itemsets) that have transaction support above minimum support → Itemsets with minimum support are called *large itemsets*, and all others small itemsets.
  - Generate from each large itemset, rules that use items from the large itemset.
    - Given a large itemset *Y*, and *X* is a subset of *Y*
    - Take the support of *Y* and divide it by the support of *X*
    - If the ratio is at least *minconf*, then $X \Rightarrow (Y - X)$ is satisfied with confidence factor c

Discovering Large Itemsets



- Algorithm for discovering large itemsets make multiple passes over the data
  - In the first pass: count the support of individual items and determine which of them are large.
  - In each subsequent pass:
    - start with a set of itemsets found to be large in the previous pass.
    - This set is used for generating new potentially large itemsets, called *candidate* itemsets
    - counts the actual support for these candidate itemsets during the pass over the data.
  - This process continues until no new large itemsets are found.

Generate rules from large itemsets

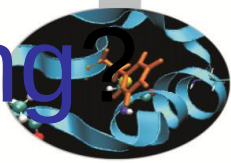


Y = {Bread, Butter, Milk}, X = {Bread, Butter}

conf = support(Y) / support(X) = {Bread, Butter, Milk} / {Bread, Butter}

if *conf* $\geq$ *minconf* then the rule {Bread, Butter} $\Rightarrow$ Milk holds

What Is Sequential Pattern Mining?



- Given a set of sequences and support threshold, find the complete set of *frequent* subsequences

A *sequence* : < (ef) (ab) (df) c b >

A *sequence database*

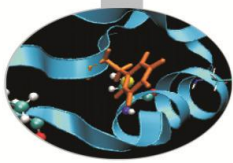
SID	sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

An element may contain a set of items. Items within an element are unordered and we list them alphabetically.

<a(bc)dc> is a *subsequence* of <a(abc)(ac)d(cf)>

Given *support threshold* *min_sup* =2, <(ab)c> is a *sequential pattern*

Applications of sequential pattern mining



- Customer shopping sequences:
  - First buy computer, then CD-ROM, and then digital camera, within 3 months.
- Medical treatments, natural disasters (e.g., earthquakes), science & eng. processes, stocks and markets, etc.
- Telephone calling patterns, Weblog click streams
- DNA sequences and gene structures