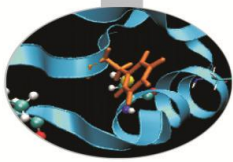
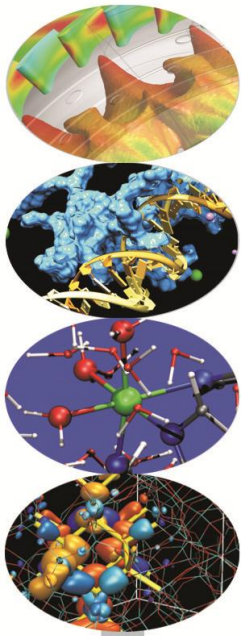
CINECA SCAI
SuperComputing Applications and Innovation



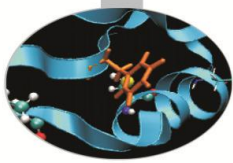
Unsupervised learning



School on Scientific Data Analytics and Visualization

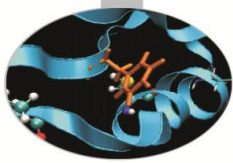
Roma, 12/06/2017

CINECA



Agenda

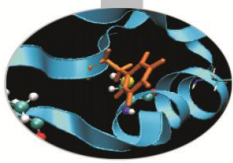
- Cluster Analysis
  - Basic concept
  - Main clustering algorithms
  - Distance measures
- Association rules and sequential patterns
  - Association Rules mining
  - Sequential Patterns mining



Cluster analysis

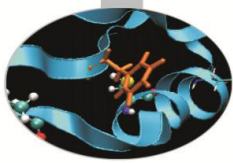
- Cluster analysis
  - no predefined classes for a training data set
  - find *similarities* between data according to characteristics underlying the data and grouping similar data objects into clusters
  - two general tasks: identify the “natural” clustering number and properly grouping objects into “sensible” clusters
- 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
- Typical applications
  - as a stand-alone tool for data exploration
  - as a precursor to other supervised learning methods

Typical applications



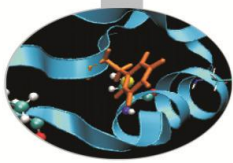
- Scientific applications
  - **Biology**: discovering genes with similar functions in DNA microarray data.
  - **Seismology**: grouping earthquake epicenters to identify dangerous zones.
  - …
- Business applications
  - **Marketing:** discovering distinct groups in customer bases (insurance, bank, retailers) to develop targeted marketing programs.
  - **Behavioral analysis**: identifying driving styles.
  - …
- Internet applications
  - **Social network analysis:** in the study of social networks, clustering may be used to recognize communities within a network.
  - **Search result ordering:** grouping of files or web pages to create relevant sets of search results.
  - …

Data representation



The data must arranged in a data matrix containing information on N objects (cases or observations ; rows of the matrix) specified by the values assigned to V variables (columns of the matrix)

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



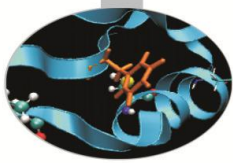
Cluster Analysis steps

- Selection of variables
- Preprocessing
- Selection of a clustering algorithm
- Selection of a distance or a similarity measure (*)
- Determination of the number of clusters (*)
- Validation of analysis

Changing one parameter may result in complete different cluster results.

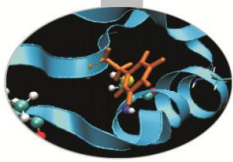
(*) if needed by the method used

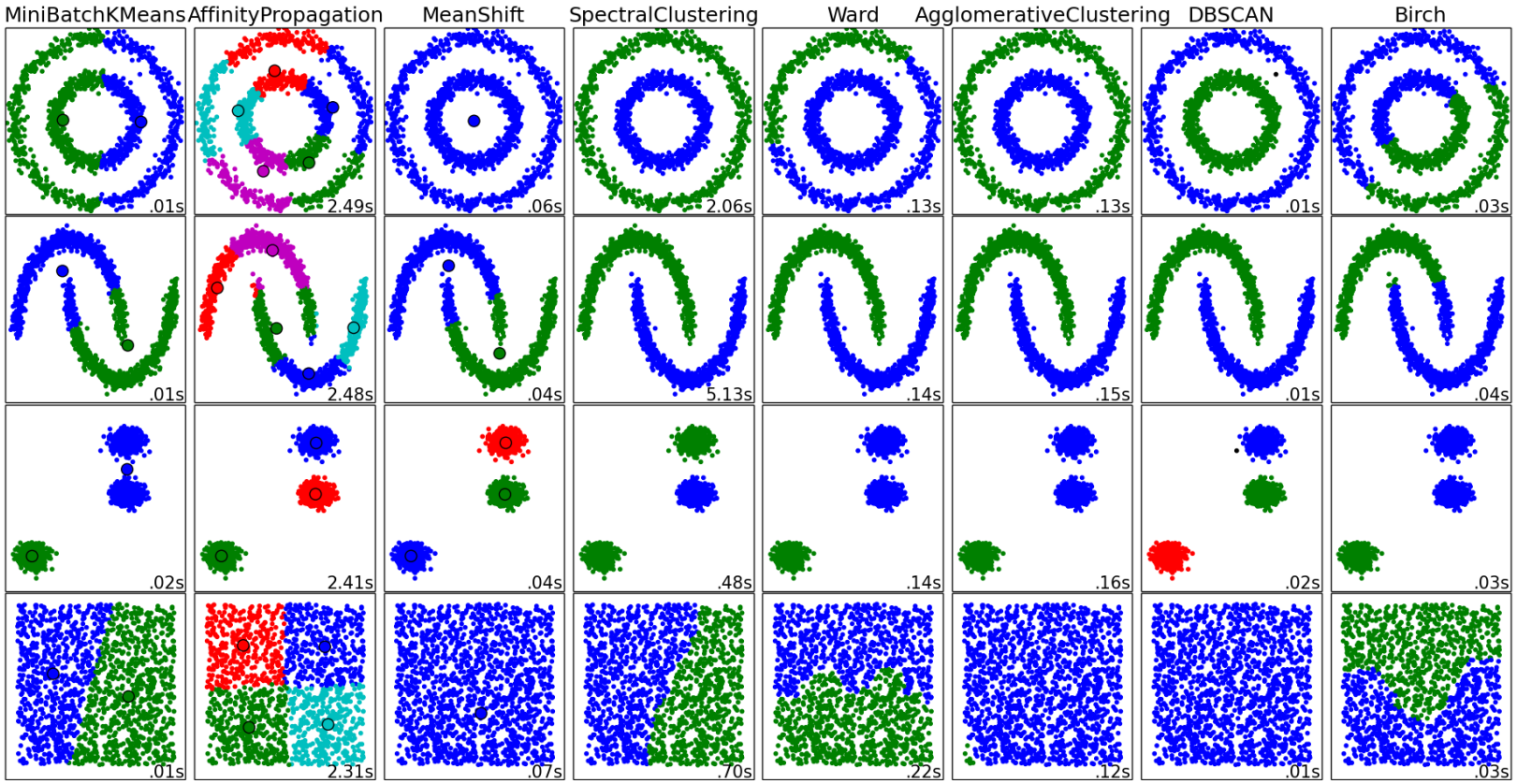
Classifications of methods



- Types of clustering (1)
  - Distance-based
  - Density-based
  - Model-based
  - Spectral
  - …
  - Combination of methods
- Types of clustering (2)
  - Partitional vs. Hierarchical
  - Complete vs. Partial
  - Exclusive vs. Fuzzy vs. Overlapping
- …

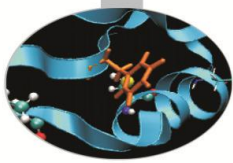
Comparison of methods





http://scikit-learn.org/stable/modules/clustering.html

Distance measure



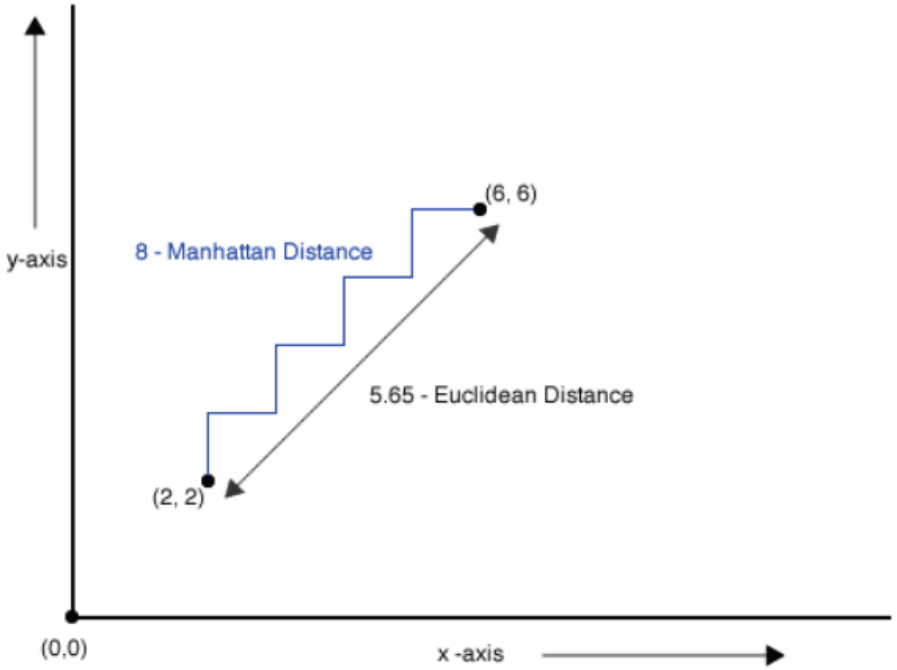
Different distance metrics give different clusters.

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)

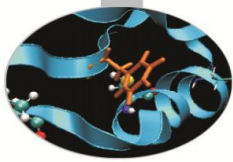
$$d(i,k) = \left[\sum_{j=1}^{d} (x_{ij} - x_{kj})^{2} \right]^{1/2}$$



Manhattan distance (city block distance)

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

Distance Measures



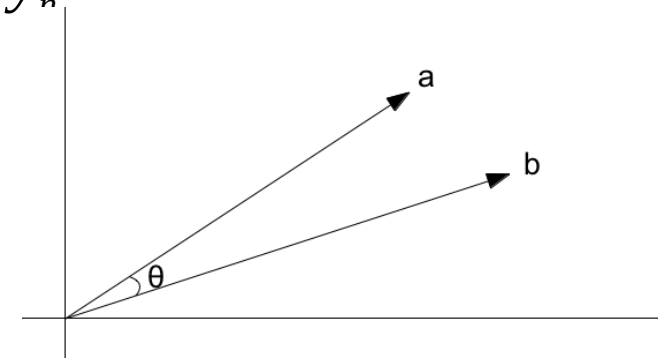
- Cosine Measure (Similarity vs. Distance)

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

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

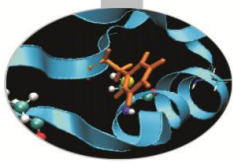
$$d(\mathbf{x},\mathbf{y}) = 1 - \cos(\mathbf{x},\mathbf{y}) = 1 - \frac{\mathbf{x} \bullet \mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}$$

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



  - Cosine similarity is a common similarity metric in text analysis.
  - It measures the smallest angle between two vectors
  - It can be turned into a pseudo distance by subtracting it from 1.0

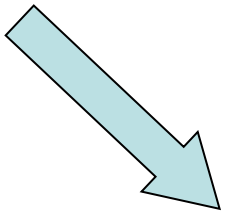
Similarity measures

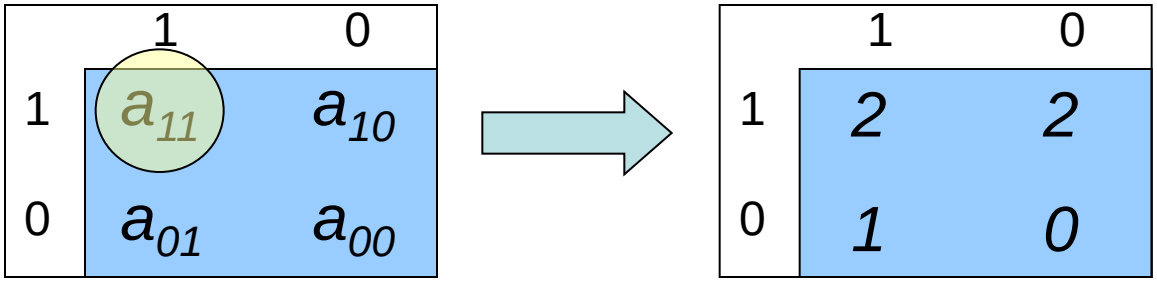


Correspondent 1's

x_k: 0 1 1 0 1

x_j: 1 1 0 1 1



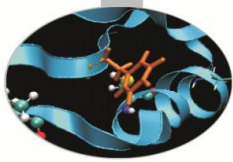


Jaccard: $d(i,k) = (a_{11}) / (a_{11} + a_{10} + a_{01})$

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

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

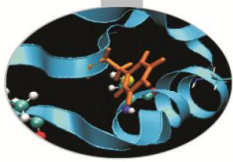
Distance-based Partitioning



- 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*: Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM: Each cluster is represented by one of the objects in the cluster

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

K-MEANS pseudo code



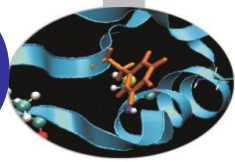
K-MEANS CLUSTERING ALGORITHM

```
1  Choose k initial centers C = {c_1, ..., c_k}
2  while stopping criterion has not been met
3      do ▷ Assignment step:
4          for i = 1, ..., N
5              do find closest center c_k ∈ C to each instance p_i
6                  assign instance p_i to cluster C_k
7          ▷ Update step:
8          for k = 1, ..., K
9              do set c_k to be the center of mass of all points in C_i
```

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)$$

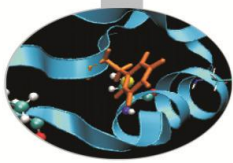
http://flashycud.com/projects/k-means-vis/src/index.html

Distance based (Hierarchical)



- Agglomerative (bottom-up):
  - Start with each observation being a single cluster.
  - Eventually all observations belong to the same cluster.
- Divisive (top-down):
  - Start with all observations belong to the same cluster.
  - Eventually each node forms a cluster on its own.
  - Could be a recursive application of k-means like algorithms
- Does not require the number of clusters *k* in advance
- Needs a termination/readout condition

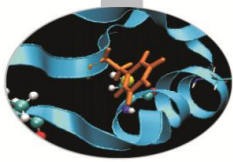
Limitation of k-means



- Works best for spherical clusters
- Cannot find odd-shaped clusters
  - [but can implicitly approximate if allowed many small clusters]

- Alternatives:
  - density-based clustering

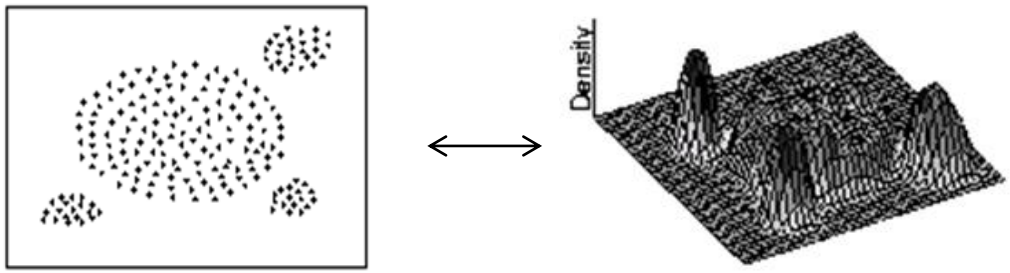
Density-based clustering



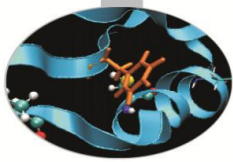
DBSCAN is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), omitting as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).

The number of clusters is determined by the algorithm.

DBSCAN does not produce a complete clustering



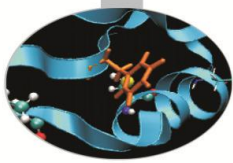
DBSCAN pseudo code



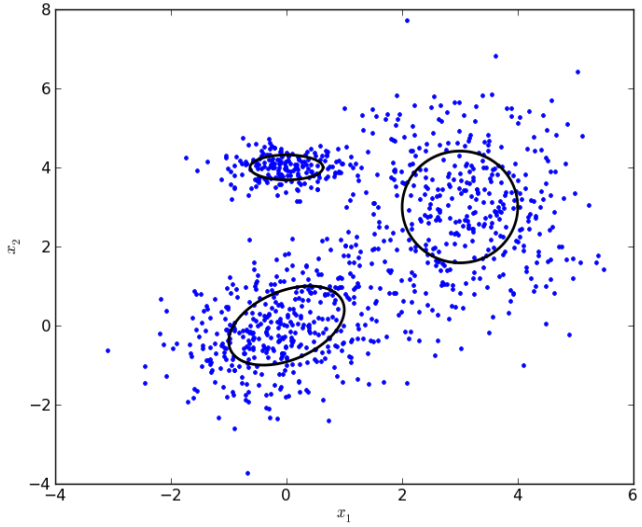
DBSCAN ALGORITHM

```
1  select an arbitrary point p
2  retrieve all points density-reachable from p
3  if p is a core point
4     then a cluster is formed
5  if p is a border point
6        do no points are density-reachable from p
7           DBSCAN visits the next point of the database
8  continue the process until all of the points have been processed
```

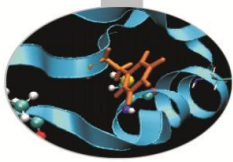
Model-based clustering



- 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)



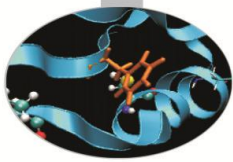
Spectral Clustering



- 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

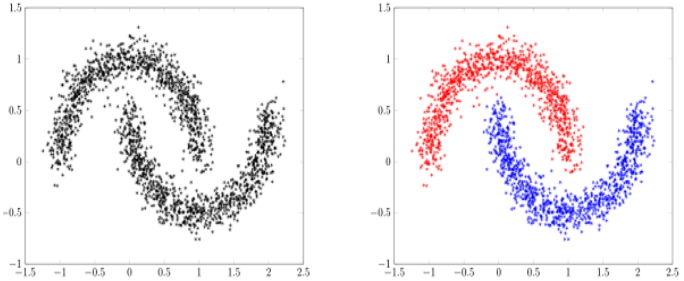


Using different methods can be useful for overcome the drawbacks of a single methods.

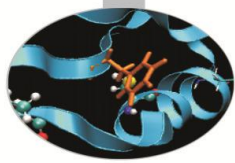
For example it is possible to generate a large number of clusers with K-means and then cluster them together using a hierarchical method.

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.

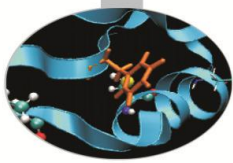


Cluster validation



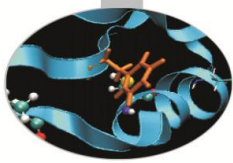
- In supervised classification, the evaluation of the resulting model is an integral part of model developing process
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters with the aim of:
  - avoiding finding patterns in noise
  - comparing clustering algorithms
  - comparing two sets of clusters
  - comparing two clusters

Cluster validation



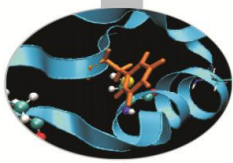
- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - e.g.: Entropy $e_i = -\sum_q \rho_{ij} \log_2 \rho_{ij}$.
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - e.g.: Sum of Squared Error (SSE) $WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$
  - Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as criteria instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

Cluster validation



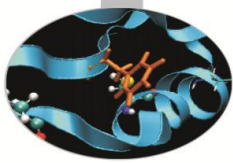
Internal measures	External measures
Gamma	Corrected Rand statistic
C Index	Jaccard coefficient
Point-Biserial	Folkes and Mallow Index
Log Likelihood	Hubert Γ statistics
Dunn's Index	Minkowski score
Tau	Purity
Tau A	van Dongen criterion
Tau C	V-measure
Somer's Gamma	Completeness
Ratio of Repetition	Homogeneity
Modified Ratio of Repetition	Variation of information
Adjusted Ratio of Clustering	Mutual information
Fagan's Index	Class-based entropy
Deviation Index	Cluster-based entropy
Z-Score Index	Precision
D Index	Recall
Silhouette coefficient	F-measure

Association Rules mining



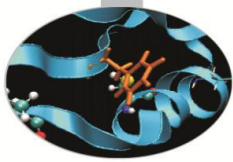
- Association rule mining is used to find objects or attributes that frequently occur together.
- For example, products that are often bought together during a shopping session (market basket analysis), or queries that tend to occur together during a session on a website's search engine.
- The unit of "togetherness" when mining association rules is called a transaction. Depending on the problem, a transaction could be a single shopping basket, a single user session on a website, or even a single customer.
- The objects that comprise a transaction are referred to as items in an itemset.

Typical applications



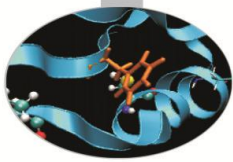
- **Market Basket Analysis**: given a database of customer transactions, where each transaction is a set of items the goal is to find groups of items which are frequently purchased together.
- **Telecommunication** (each customer is a transaction containing the set of phone calls)
- **Credit Cards/ Banking Services** (each card/account is a transaction containing the set of customer's payments)
- **Medical Treatments** (each patient is represented as a transaction containing the ordered set of diseases)
- **Basketball-Game Analysis** (each game is represented as a transaction containing the ordered set of ball passes)

Definitions



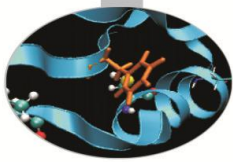
- A set of items is referred as an **itemset**. A itemset that contains k items is a k-itemset.
- The **support of an itemset** X is the percentage of transactions in the transaction database D that contain X.
- The **support of the rule** $X \Rightarrow Y$ in the transaction database D is the support of the items set $X \cup Y$ in D.
- The **confidence of the rule** $X \Rightarrow Y$ in the transaction database D is the ratio of the number of transactions in D that contain $X \cup Y$ to the number of transactions that contain X in D.

Market Basket Analysis



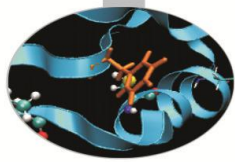
- Basket data consist of collection of transaction date and items bought in a transaction
- Retail organizations interested in generating qualified decisions and strategy based on analysis of transaction data
  - what to put on sale, how to place merchandise on shelves for maximizing profit, customer 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 rule discovery problem



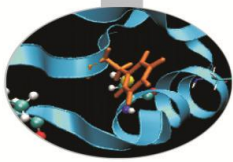
- Two sub-problems in discovering all association rules:
  - Find all sets of items (itemsets) that have transaction support above minimum support → Itemsets with minimum support are called *large itemsets*, and all others small itemsets.
  - Generate from each large itemset, rules that use items from 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

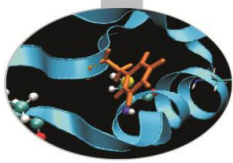


Y = {Bread, Butter, Milk}, X = {Bread, Butter}

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

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

Lift

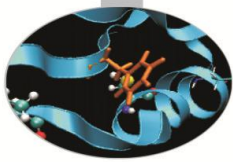


- The lift measure indicates the departure from independence of X and Y. The lift of $X \Rightarrow Y$ is :

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{p(Y)} = \frac{\frac{p(X \wedge Y)}{p(X)}}{p(Y)} = \frac{p(X \wedge Y)}{p(X)p(Y)}$$

- But, the lift measure is symmetric; i.e., it does not take into account the direction of implications!

Sequential Pattern Mining



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

A *sequence* : <(ef)(ab)(df)cb>

A *sequence database*

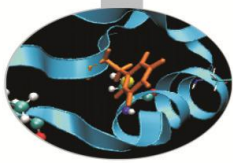
SID	sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

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)>

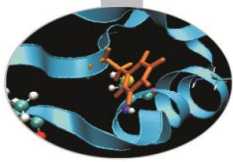
Given *support threshold* *min_sup* =2, <(ab)c> is a *sequential pattern*

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

References



- An Introduction to Statistical Learning
  - R. Tibshirani e T. Hastie
- The Elements of Statistical Learning
  - J.H. Friedman, R. Tibshirani e T.Hastie
- Analisi dei dati e Data Mining
  - A. Azzalini e B. Scarpa