Grid Based Methods

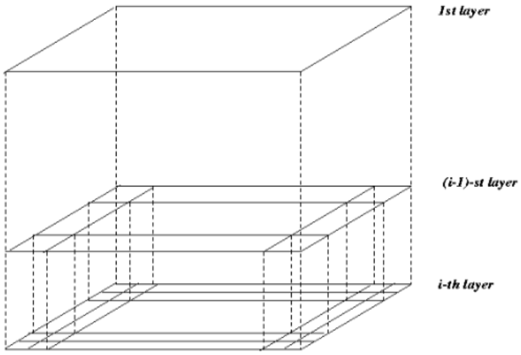
Grid-Based Clustering method uses a multi-resolution grid data structure.

Several interesting methods

- **STING** (a **ST**atistical **IN**formation **G**rid 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

STING was proposed by Wang, Yang, and Muntz (VLDB'97). In this method, the spatial area is divided into rectangular cells.There are several levels of cells corresponding to different levels of resolution.For each cell, the high level is partitioned into several smaller cells in the next lower level. The statistical info of each cell is calculated and stored beforehand and is used to answer queries.



The 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.

Then using a top-down approach we need to answer spatial data queries. Then start from a pre-selected layer—typically with a small number of cells. For each cell in the current level compute the confidence interval. Now remove the irrelevant cells from further consideration. When finishing examining the current layer, proceed to the next lower level.
Repeat this process until the bottom layer is reached.

Advantages:

It is 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

It was proposed by Sheikholeslami, Chatterjee, and Zhang (VLDB'98).

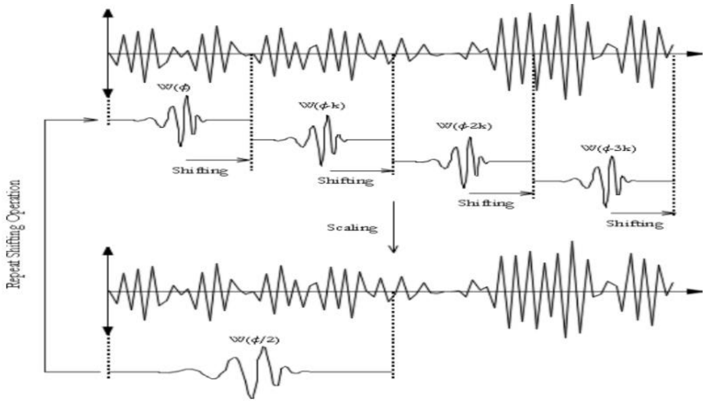
It is a multi-resolution clustering approach which applies wavelet transform to the feature space

- A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-band.

It can be both grid-based and density-based method.

Input parameters:

- No of grid cells for each dimension
- The wavelet, and the no of applications of wavelet transform.



How to apply the wavelet transform to find clusters

- It summaries the data by imposing a multidimensional grid structure onto data space.
- These multidimensional spatial data objects are represented in an n-dimensional feature space.
- Now apply wavelet transform on feature space to find the dense regions in the feature space.
- Then apply wavelet transform multiple times which results in clusters at different scales from fine to coarse.

Why is wavelet transformation useful for clustering

- It uses hat-shape filters to emphasize region where points cluster, but simultaneously to suppress weaker information in their boundary.
- It is an effective removal method for outliers.
- It is of Multi-resolution method.
- It is cost-efficiency.

Major features:

- The time complexity of this method is O(N).
- It detects arbitrary shaped clusters at different scales.
- It is not sensitive to noise, not sensitive to input order.
- It only applicable to low dimensional data.

CLIQUE - Clustering In QUEst

It was proposed by Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98). It is based on automatically identifying the subspaces of high dimensional data space that allow better clustering than original space.
CLIQUE can be considered as both density-based and grid-based:

- It partitions each dimension into the same number of equal-length intervals.
- It partitions an m-dimensional data space into non-overlapping rectangular units.
- A unit is dense if the fraction of the total data points contained in the unit exceeds the input model parameter.
- A cluster is a maximal set of connected dense units within a subspace.

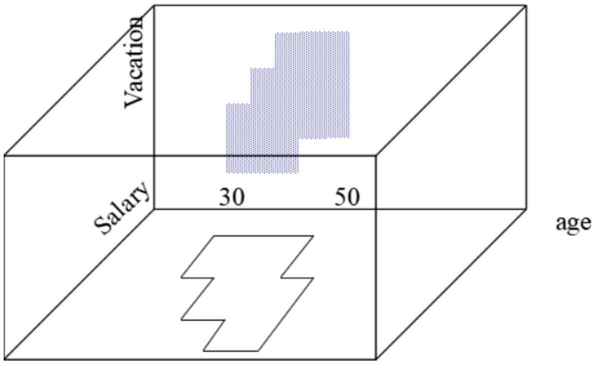
Partition the data space and find the number of points that lie inside each cell of the partition. Identify the subspaces that contain clusters using the Apriori principle.

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.



Advantages

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.

Disadvantages

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

Model-Based Clustering

Model-based clustering method is an attempt to optimize the fit between the data and some mathematical models. It is the Statistical and AI approach. Model-based clustering works on the intuition that gene expression data originates from a finite mixture of underlying probability distributions .Each cluster corresponds to a different distribution , and these distributions are assumed to be Gaussians. The parameters of each distribution (i.e., cluster) are estimated by maximizing the likelihood of the expression data (Hogg and Craig 1994). The *k*-means clustering method is a special case of model-based clustering, where *all* the distributions are assumed to be Gaussians with equal variance.

- Randomly generate the parameters (the parameters would be the mean and standard deviation or covariance matrix) describing each probability distribution (i.e., cluster)
- Repeat until the parameters of each distribution converge
  - For each gene, estimate the probability that the gene's expression pattern was generated from each of the distributions.
  - For each distribution, estimate the parameters of the distribution to maximize the likelihood of the expression data given the probability that each gene was generated from the distribution.
- Assign each gene to the distribution which generates the gene's expression profile with maximum probability

Model-based clustering has the advantage of providing the probability that each gene belongs in each cluster.

However, model-based clustering operates under the assumption that expression data comes from particular probability distributions, which may not be a reasonable assumption for many microarray data sets.

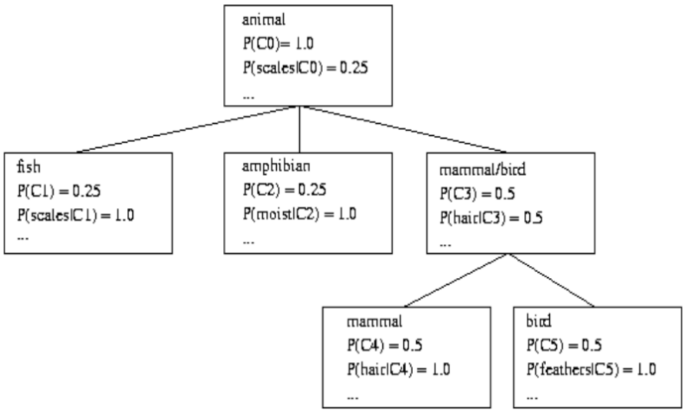
Conceptual clustering

Conceptual clustering is a form of clustering in machine learning. It produces a classification scheme for a set of unlabeled objects and finds characteristic description for each concept (class).

COBWEB (Fisher'87)

COBWEB is a popular a simple method of incremental conceptual learning.It 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.

Classification Tree



Limitations of COBWEB

The assumption that the attributes are independent of each other is often too strong because correlation may exist. It is not suitable for clustering large database data – skewed tree and expensive probability distributions. Some of the other methods alike COBWEB are:

CLASSIT

- It is an extension of COBWEB for incremental clustering of continuous data.
- It suffers similar problems as COBWEB.

AutoClass (Cheeseman and Stutz, 1996)

- It uses Bayesian statistical analysis to estimate the number of clusters.
- It has been popular in the industry.

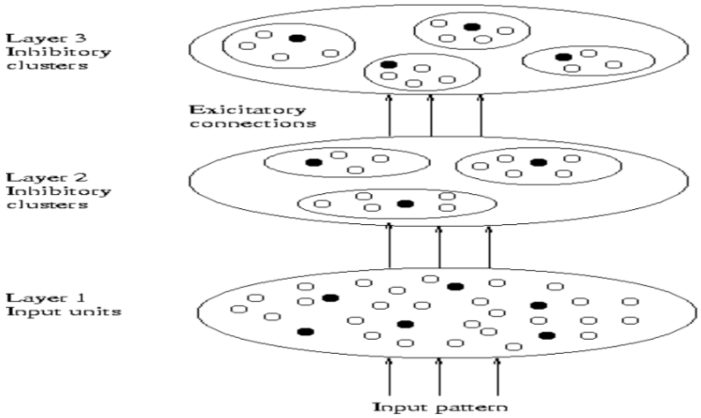
Other Model-Based Clustering Methods

<u>Neural network approaches</u>

- It represents each cluster as an exemplar, acting as a “prototype” of the cluster.
- Then new objects are distributed to the cluster whose exemplar is the most similar according to some distance measure.

Competitive learning

- It involves a hierarchical architecture of several units (neurons).
- Neurons compete in a “winner-takes-all” fashion for the object currently being presented.



Self-Organizing Feature Maps

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-D or 3-D space.