UNIT IV CLASSIFICATION AND CLUSTERING 10

Decision Tree Induction - Bayesian Classification – Rule Based Classification – Classification by Back propagation – Support Vector Machines – Associative Classification – Lazy Learners – Other Classification Methods – Clustering techniques – , Partitioning methods- k-means- Hierarchical Methods – distance based agglomerative and divisible clustering, Density-Based Methods – expectation maximization -Grid Based Methods – Model-Based Clustering Methods – Constraint – Based Cluster Analysis – Outlier Analysis

Clustering techniques - Grid Based Methods

- The grid-based clustering methods **use a multi-resolution grid data structure**. It quantizes the object areas into a finite number of cells that form a grid structure on which all of the operations for clustering are implemented.

- The **benefit of the method is its quick processing time**, which is generally independent of the number of data objects, still dependent on only the multiple cells in each dimension in the quantized space.

- An instance of the grid-based approach involves **STING**, which explores **statistical data stored in the grid cells**, WaveCluster, which clusters objects using a wavelet transform approach, and **CLIQUE**, which defines a **grid-and density-based approach** for clustering in high-dimensional data space.

STING

- **STING** (Statistical Information Grid Clustering Algorithm) and **OPTICS** (**Ordering Point To Identify Clustering Structure Clustering Algorithm**) are clustering algorithms used in Unsupervised Learning.
- They are machine learning techniques which are used to club the given input data points into clusters or groups on the basis of their attributes.
- **STING** is grid-based clustering algorithm while **OPTICS** is a density-based clustering algorithm.

S.No.	STING	OPTICS
1.	STING is abbreviation for **Statistical Information Grid**	OPTICS is abbreviation for **Ordering Point To Identify Clustering Structure**
2.	It is **grid based** clustering algorithm	It is **density based** clustering algorithm
3.	It concerns not with data points but with the value space that surrounds the data points.	It searches the data space for areas of varied **density data points** in the data space.
4.	It uses multi-dimensional grid data structure that quantizes space into a finite number of cells.	It is an extension to Density Based spatial clustering of applications with noise.
5.	The following are the properties of STING clustering algorithm: •Spatial area is divided into rectangular cells. •Several level of cells at different levels of resolution. •High level cell is partitioned into several low level cells. •Statistically attributes are stored in cell for instance Mean, Maximum, Minimum are some of the statistical measures which are used. •Statistical information is calculated for each cell and the types of distribution calculated are normal and exponential.	The following are the properties of OPTICS clustering algorithm: •It is an extension of DBSCAN, which takes the responsibility of parameters that can lead to discovery of unacceptable clusters. •Core distance is the smallest point that make a point core. •Two important parameters are required for OPTICS: epsilon(“eps) and minimum points(“MinPts). •The parameter eps defines the radius of neighborhood around a point P. The parameter MinPts is the minimum no. of neighbors within “eps”radius. •Density = No. of points within a specified radius r(eps)
		It has relatively more computational

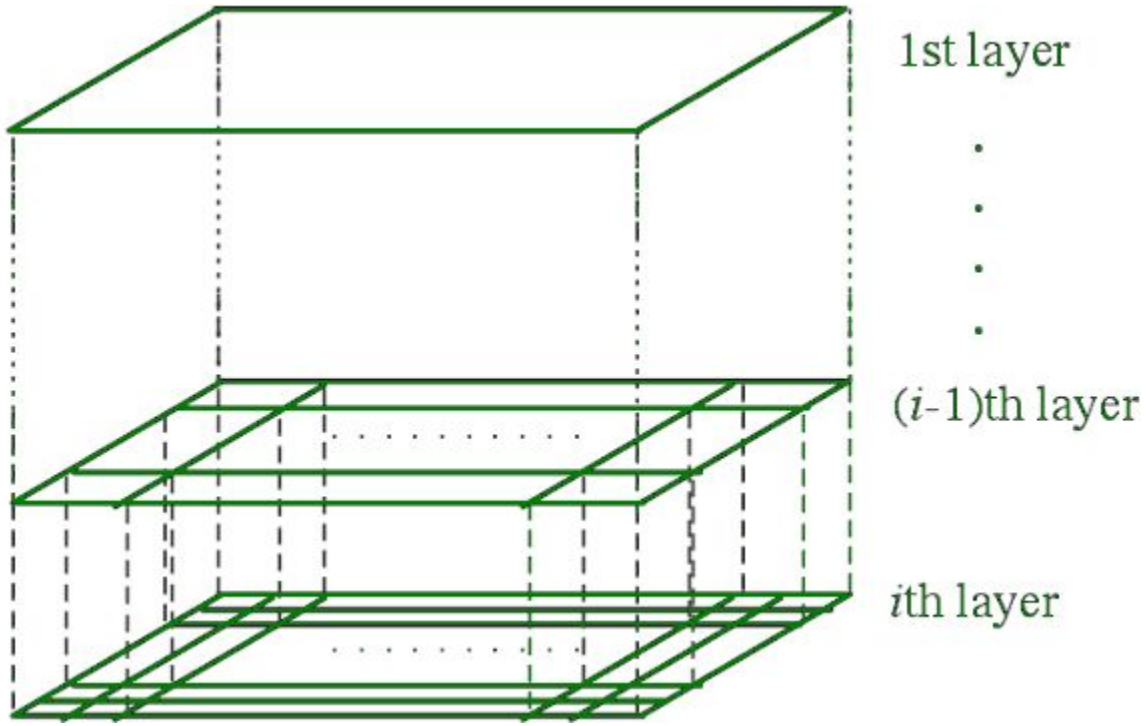
STING Algorithm:

1. **Determine a layer**, to begin with.
2. **For each cell of this layer, we calculate the confidence interval** (or estimated range) of probability that this cell is relevant to the query.
3. **From the interval calculate above, we label the cell as relevant or not relevant**.
4. **If this is the bottom layer, then end the process**.
5. We go down the hierarchy structure by one level. Go to Step 2 for those levels that form the relevant cells of the higher-level layer.

STING Hierarchy Diagram :

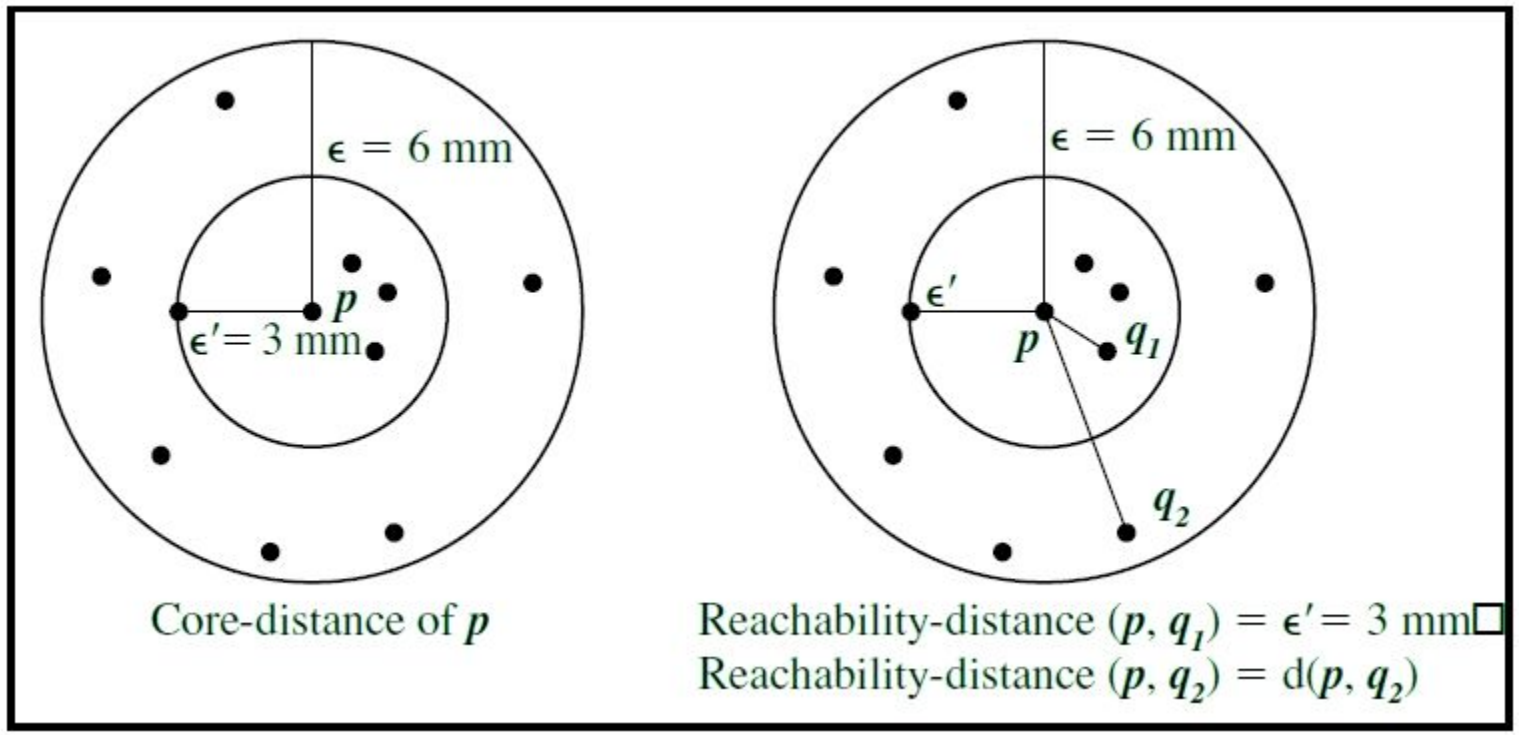
1st level (top level) could have only one cell.

A cell of (i-1)th level corresponds to 4 cells of ith level.



OPTICS Algorithm:

- **Core distance of a point P is the smallest distance such that the neighborhood of P has atleast minPts points.**
  Reachability distance of p from q1 is the core distance (ε’).
  Reachability distance of p from q2 is the euclidean distance between p and q2.



Core-distance of p

Reachability-distance $(p, q_1) = \epsilon' = 3$ mm

Reachability-distance $(p, q_2) = d(p, q_2)$