Friday
12-06-2020

DWDM Assignment-4

UNIT-4

N.V.S.K. kalyani
17Y51A0584
3rd B.Tech CSE-C

2 Marks

1A) cluster analysis:-

The process of collecting homogeneous data objects within the same group called cluster & heterogeneous to the objects in the other group is called as cluster analysis.

2A) Hierarchical clusters:-

- A Hierarchical clustering method works by grouping data projects into tree of the clusters. The processing of the hierarchical clustering is slower than partitional clusters.
- Input parameters are not required.
- They do not need assumptions other than a similarity measures.

partitioning clustering:-

- partitioning cluster method divides the given data base of n objects into m partitions such as $m <= 0$ where each partition is a clustering.
- They require the certain Input parameters to start processing to start processing is faster than hierarchical cluster.

3A) STING:- Statistical Information grid is a grid based method. The region of the space is divided into the cells of the rectangular shapes.

- These cells are divided into the several corresponding to the resolution and the cells at the higher level further partitioned into the cells at next

4A) DENCLUE:- means DENsity-based CLUstEring) is a clustering method based on a set of density distribution functions.

- The method is built on the following idea that the influence of each data point can be formally modeled using a mathematical function, called an influence function, which describes the impact of a data point within its neighborhood.

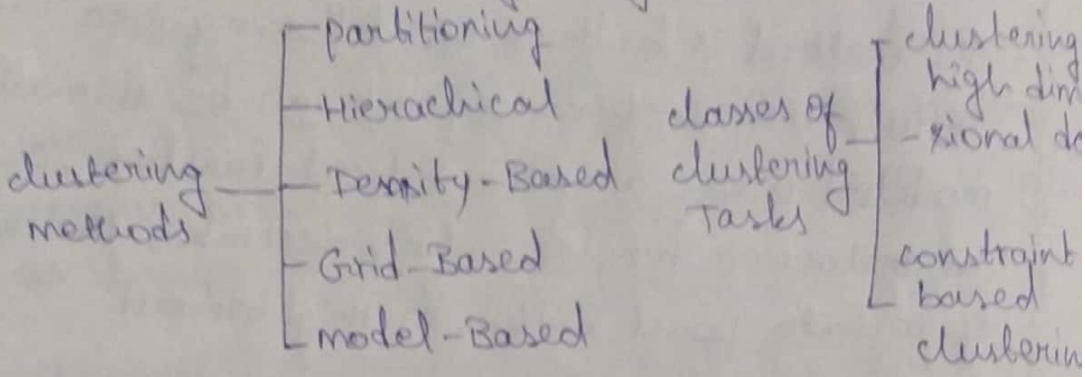
5A) Types of clustering method:-

The important clustering methods can be grouped into the following categories:

1. Model based methods
2. Grid based methods
3. Density based methods
4. partitioning methods
5. Hierarchical methods

10 Marks

10M) The major categorization clustering methods can classified into following categories.



- partitioning method:- constructs various partitions and then evaluate them by some criterion.
  Ex:- Minimizing the sum of square errors.
- Hierachical method:- create a hierarchical decompo-sition of the set of data using some criterion.
  This method can be classified based on
  * Agglomerative approach (Bottom up)
  * Divisive approach (Top Down)
  Typical methods are BRICH, ROCK, chameleon and so
- Density-Based method:- It is based on connec-tivity and density functions. Typical methods are DBSCAN, OPTICS, DENCLUE, etc.
- Grid-Based method:- It is based on multiple-level granularity structure. Typical methods are STING wave cluster and clique etc.,
- model-Based methods:- This model is hypothesized for each of the clusters and tries to find the best fit of that model to each other. Typical methods are EM, SOM and COBWEB etc.

classes of clustering tasks are as follows

clustering High-dimensional data :- It is a particularly important task in cluster analysis because many application require the analysis of objects containing a large no. of features or dimensions.

constraint-Based clustering :- It is a clustering approach that performs clustering by incorporation of user-specified or application-oriented constraints. It focus on mainly

* Spatial clustering
* Semi-supervised clustering.

b) Partitioning method:-

Constructs various partitions and then evaluate that by some criterion and following diagram various typical methods, partitioning methods find sphere-shaped clusters.

1. classical partitioning methods:- The most common used are k-means & k-medoid methods.

Each cluster is represented by the center of the cluster. minimum sum of the squared distance can be calculated as (square error criterion).

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

where E is the sum of the square error for all objects in the data set ; p is the point in space representing a given object ; and m_i is the mean of cluster C_i.

The k-mean is efficient for large data sets but sensitive to outliers.

* The Expectation Maximization algorithm extends the k-means paradigm in a different way.

Representative object-Based Technique :-

The k-medoids method instead of taking the mean value of the objects in a cluster as a represent point pick actual objects to represent the cluster, using one representative object per cluster.

The partitioning method is then performed based on the principle of minimizing the sum of the dissimilarties b/w each other and its corresponding reference point. That is an absolute-error criterion is used, defined as

$$E = \sum_{j=1}^{k} \sum_{P \in C_j} |P - O_j|$$

2. partitioning methods in large database: from k-medoids to CLARANS:

CLARA (clustering large Applications) uses a Sampling based method to deal with large data sets.

CLARANS (clustering large Application based upon RANdomized search) was proposed to improve the quality and the scalability of CLARA.

Hierarchical Methods :- works by grouping data objects into of clusters. These methods are classified as follows:

* Agglomerative and divise Hierarchical clustering

2A) Model-Based clustering methods attempt to optimize blw the given data and some mathematical method. These examples of model-based clustering are as follows:

* Expectation-maximization - presents an Extension of the k-means positioning algorithm.
* Conceptual clustering.
* Neural Network Approach

1. Expectation-maximization :-

A popular iterative refinement algorithm. An Extension to k-means.

* Assign each object to a cluster according to a weight.
* New means are computed based on weighted measures.

The process can be follows:

* starts with an initial estimate of the parameter vector.
* Mesatively resocores the patterns against the mixture density produced by the parameter vector
* The resecored patterns are used to update the parameter update.
* patterns belonging to the same cluster, if they are placed by their scores in a particular component

The EM Algorithm is as follows:

* Initialy, random assign k cluster contest.

* Iteratively refine the clusters based on 2 steps
  -> Exception step: Assign each object x_i to cluster C_k with probability as

$$P(x_i \in C_k) = p(c_k|x_i) = \frac{P(C_k)P(x_i|c_k)}{P(x_i)}$$

  where $p(x_i|C_k) = N(m_k, E_k(x_i))$ Follows the normal distribution around mean, m_k, with expectation, E_k

  -> Maximization step: Estimation of model parameters

$$m_k = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i P(x_i \in C_k)}{\sum_i P(x_i \in C_i)}$$

This step is the "maximization" of the likelyhood of the distribution.

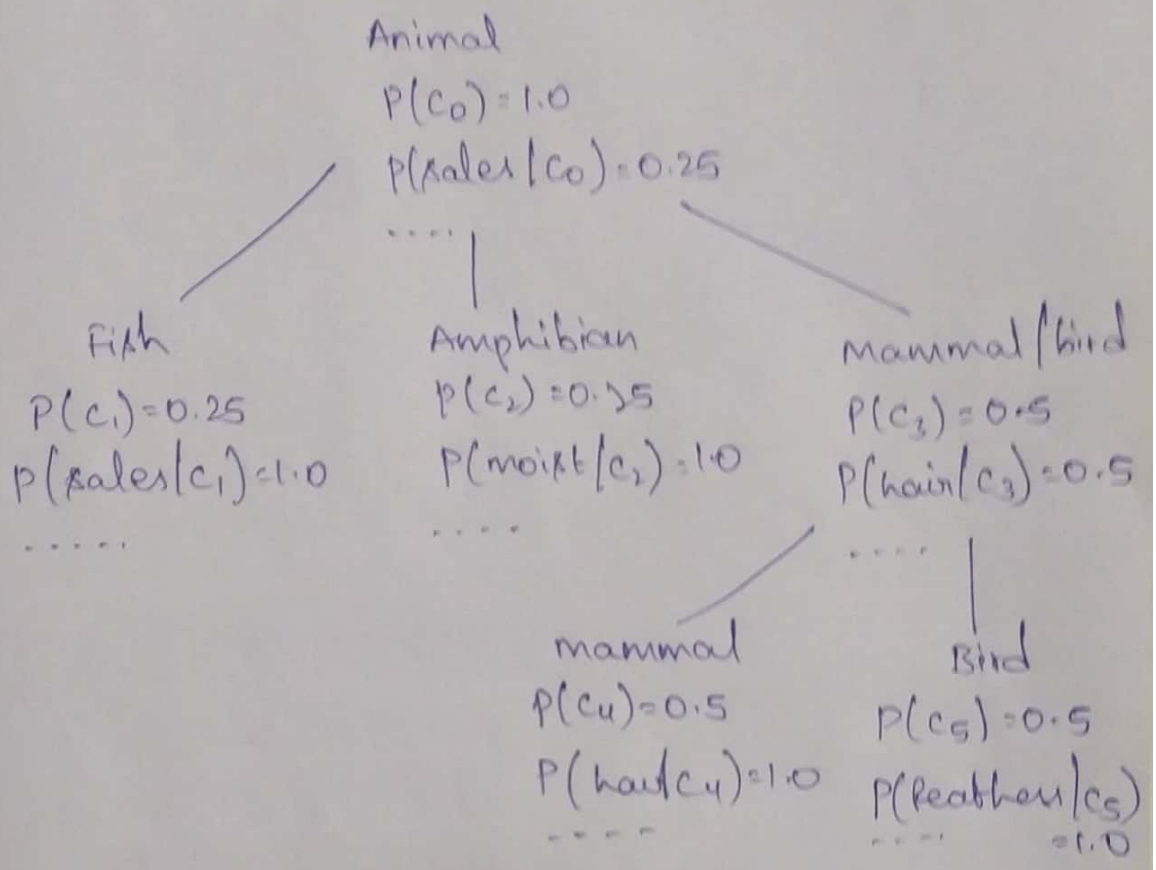
2. Conceptual clustering :-

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

COBWEB is

* popular simple method of incremental conceptual learning.
* creates a hierarchical clustering in the form of classification tree.
* each node refers to a concept and contains a probalistic of the concept.

$$\frac{\sum_{k=1}^{n} P(C_k)\left[\sum_i \sum_j P(A_i = V_{ij}(C_k)^2 - \sum_i \sum_j P(A_i = V_{ij})^2\right]}{n}$$



Limitations of COBWEB of as follows :-

* The assumption that the attributes are Independent, of each other is often to strong because correlation may exist.
* Not suitable for clustering large database data- skewed tree and Expensive.

CLASSIT : It is an Extension of COBWEB for Incremental clustering of continous data suffers similar problem as COBWEB.

Auto class : uses Bayesian statistical analysis to estimate the no.of clusters, popular in Industry.