CLUSTERING

Overview

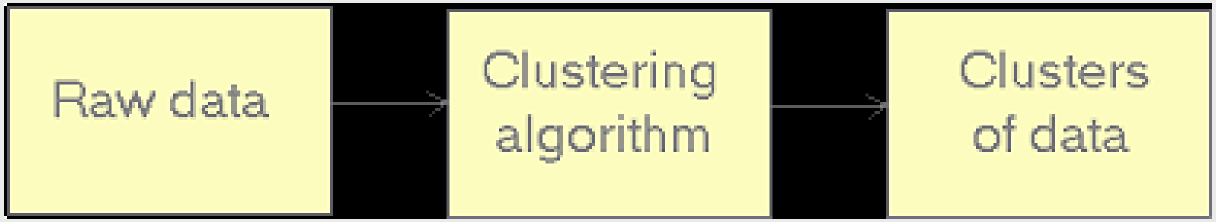
- Definition of Clustering
- Existing clustering methods
- Hierarchical clustering
- K-means
- K-means for large datasets.

Definition

- Clustering can be considered the most important *unsupervised learning* technique; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

- Clustering is “the process of organizing objects into groups whose members are similar in some way”.

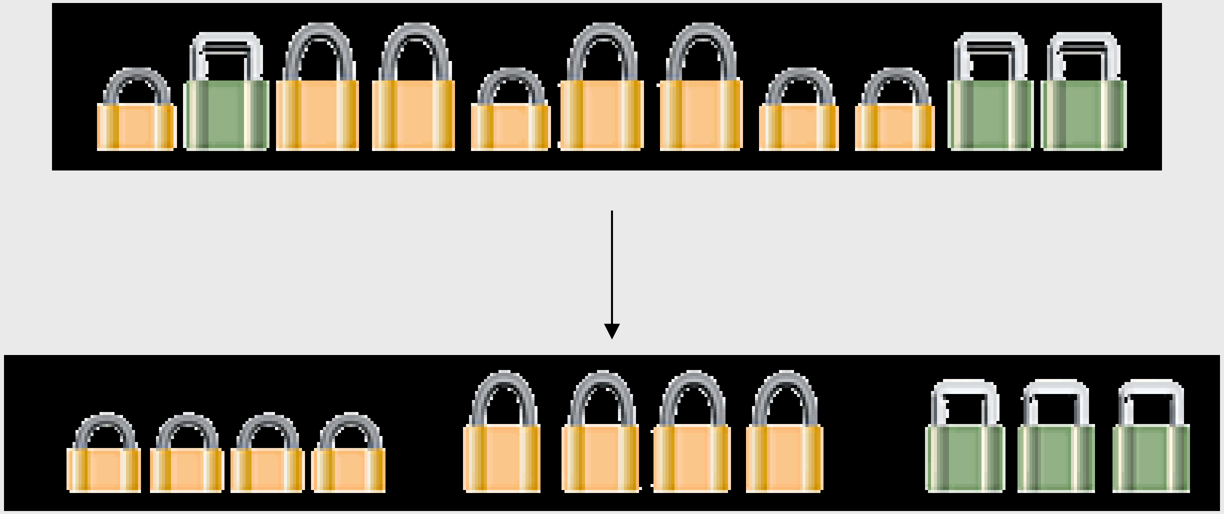
- A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.



Raw data

Clustering algorithm

Clusters of data



Why clustering?

A few good reasons ...

- Simplifications
- Pattern detection
- Useful in data concept construction
- Unsupervised learning process

Where to use clustering?

- Data mining

  -intermediate step-classification or
  outlier analysis.
- Collaborative filtering

  -summarization of like minded users
- Customer Segmentation

  -recommendations for users.
- Data summarization

  -dimensionality reduction;easier to process and interpret
- Dynamic trend Detection

  -clustered into streams to identify pattern changes

Where to use clustering?

- Multimedia data analysis

  -music,documents,video or mix

  -Determine similar segments

- Biological data analysis

  -identify gene patterns

  -structured or sequenced

- Social network analysis

  -community network summarization

Constraints to consider

- Type of attributes in data
- Scalability to larger dataset
- Ability to work with irregular data
- Time cost
- complexity
- Data order dependency
- Result presentation

Measuring Similarity

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric: $d(i, j)$
- Some distance calculation methods are The Euclidean distance,Jaccard distance and the Cosine distance.
- Distance metric constraints:
  - Triangle inequality:d(xi,xk)<=d(xi,xj)+d(xi,xk)
  - d(xi,xk)=0 => xi=xj.

Curse of Dimensionality

- Coined by Richard Bellman.
- Associated with the huge size of datasets.
- Data points become sparse-”lost in space”
- Clustering requires objects to be similar or close to each other. They are usually represented in the form of histograms where peaks represent the distance between the points-Dimensionality reduces that distance.

I.e

$$\lim \frac{(\text{max dist}-\text{min dist})}{\text{min dist}} = 0 \text{ as } d\text{->}\infty$$

Major Existing clustering methods

- Distance-based
- Hierarchical
- Partitioning
- Probabilistic

Hierarchical clustering

Agglomerative (bottom up)

1. start with 1 point (singleton)
2. recursively add two or more appropriate clusters
3. Stop when k number of clusters is achieved.

Divisive (top down)

1. Start with a big cluster
2. Recursively divide into smaller clusters
3. Stop when k number of clusters is achieved.

Hierarchical clustering types

- Single linkage clustering

  Distance b/w clusters is the **shortest** distance b/w any one member of one cluster to any member of other cluster.

- Complete link clustering

  Distance b/w clusters is the **longest** distance b/w any one member of one cluster to any member of other cluster.

Hierarchical clustering types

- Average link clustering

  Distance b/w clusters is the **average** of all distances between every pair of points from both clusters.

- Centroid link clustering

  Distance b/w clusters is the distance beyween the two means of the data points of the clusters.

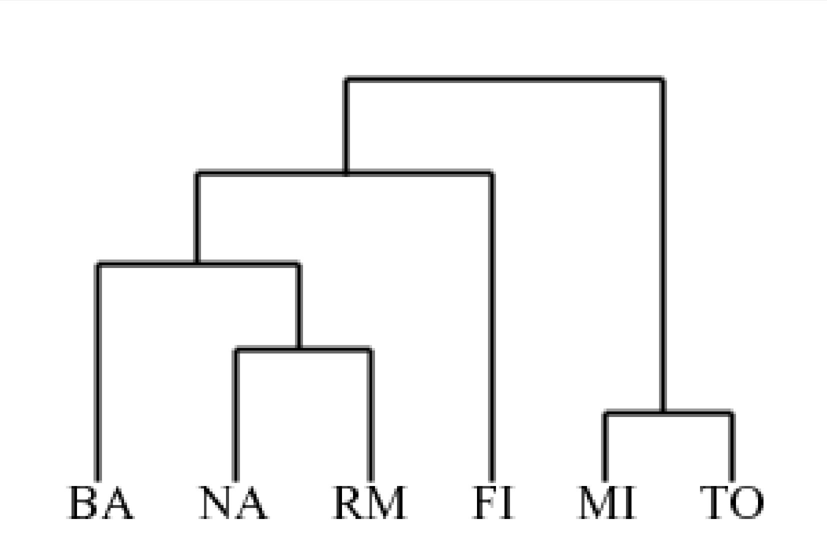
Stopping conditions:Hierarchical clustering

- Predetermined number of clusters is reached.
- Across splits or merges-little noticeable change.
- Max distance between any 2 points exceeds threshold.

Disadvantages:

- Inability to scale well.
- Undoing not possible

EXAMPLE OF A DENDROGRAM



K-mean algorithm

- It accepts the **number of clusters** to group data into, and the **dataset** to cluster as input values.
- It then creates the first **K initial clusters** (K= number of clusters needed) from the dataset by choosing K rows of data randomly from the dataset.
- ***For Example***, if there are 10,000 rows of data in the dataset and 3 clusters need to be formed, then the first **K=3 initial clusters** will be created by selecting 3 records randomly from the dataset as the initial clusters. Each of the 3 initial clusters formed will have just one row of data.

- The K-Means algorithm calculates the **Arithmetic Mean** of each cluster formed in the dataset. The Arithmetic Mean of a cluster is the mean of all the individual records in the cluster. In each of the first K initial clusters, their is only one record. The Arithmetic Mean of a cluster with one record is the set of values that make up that record.
- K-Means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset. The arithmetic mean of a cluster is the arithmetic mean of all the records in that cluster.

- This new arithmetic mean becomes the center of this new cluster. Following the same procedure, new **cluster centers** are formed for all the existing clusters.K-Means re-assigns each record in the dataset to **only one** of the new clusters formed.
- A record or data point is assigned to the **nearest cluster** (the cluster which it is most similar to) using a measure of distance or similarity .The preceding steps are repeated until **stable clusters** are formed and the K-Means clustering procedure is completed.
- Stable clusters are formed when new iterations or repetitions of the K-Means clustering algorithm does not create new clusters as the cluster center or Arithmetic Mean of each cluster formed is the same as the old cluster center.

K-means for large datasets

- Designed for datasets that dont fit into the main memory.Output is just the cluster centroids and not the clusters per say.
- The BFR algorithm assumes data points from n-dimensional Euclidean space.
- Quality of cluster determined by the variance of the points within the cluster.
- Stores parameters such as N(number of points in cluster),i-Sum of the i_{th} coordinates of the points,SUM-sum of squares of i_{th} coordinates.

Reference

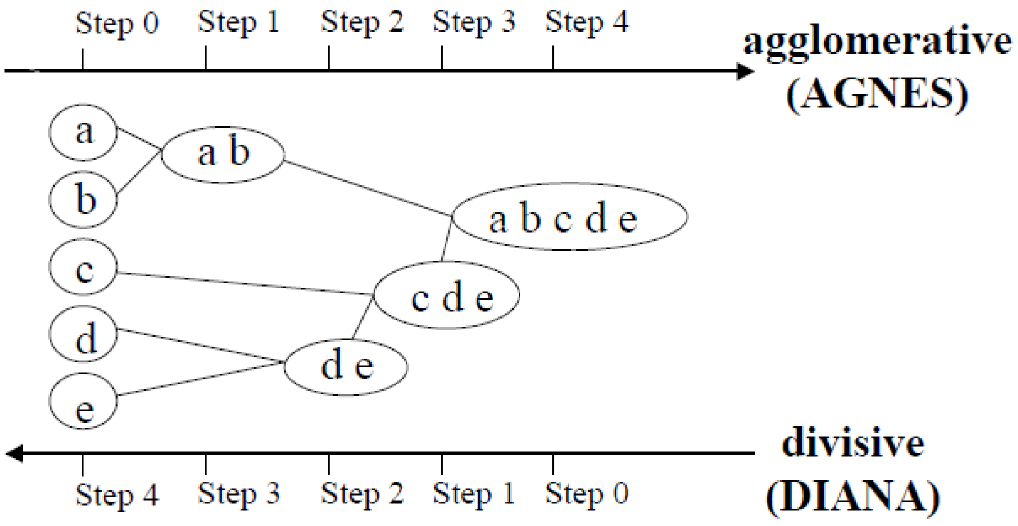
- Dr. M.H. Dunham - http://engr.smu.edu/~mhd/dmbook/part2.ppt.
- Dr. Lee, Sin-Min – San Jose State University
- Mu-Yu Lu, SJSU
- Database System Concepts, Silberschatz, Korth, Sudarshan

Five Categories of Clustering Methods

- Partitioning algorithms: Construct various partitions and then evaluate them by some criterion
- Hierarchical algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Density-based algorithms: based on connectivity and density functions
- Grid-based algorithms: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition



Step 0 Step 1 Step 2 Step 3 Step 4

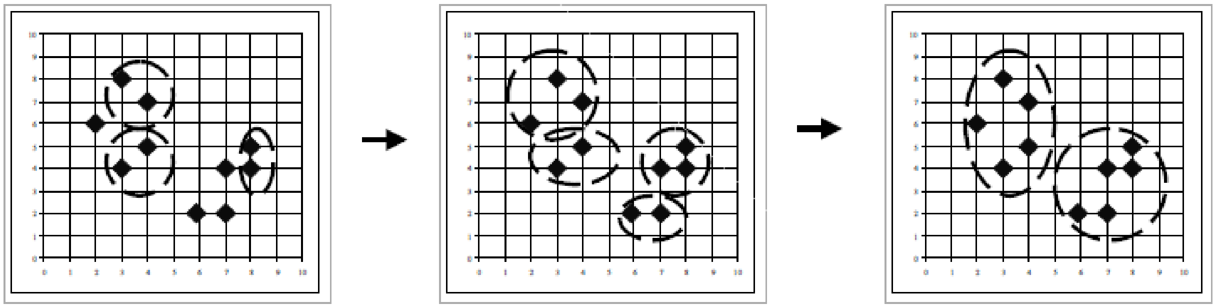
agglomerative (AGNES)

Step 4 Step 3 Step 2 Step 1 Step 0

divisive (DIANA)

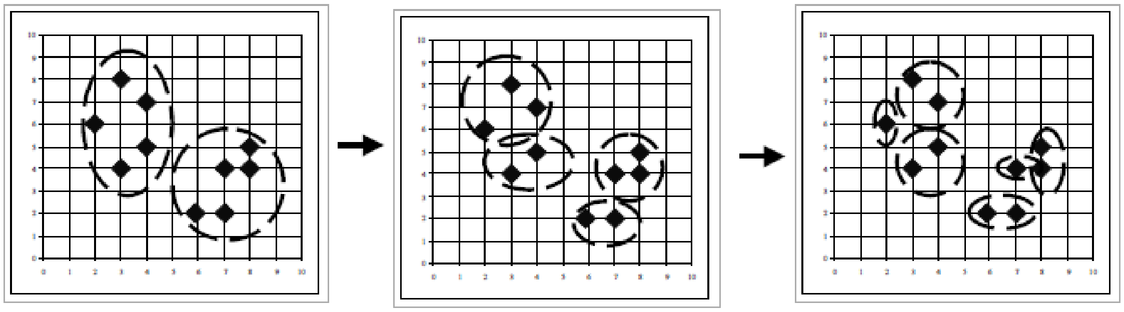
AGNES (Agglomerative Nesting)

- Agglomerative, Bottom-up approach
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster

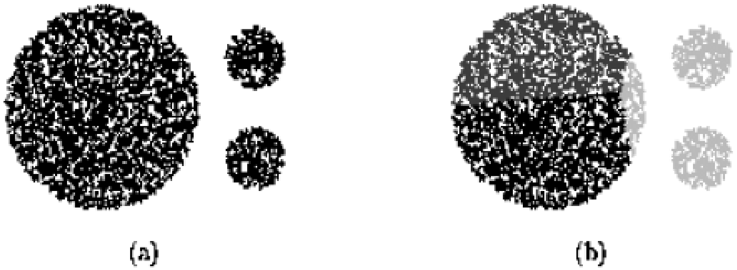


DIANA (Divisive Analysis)

- Top-down approach
- Inverse order of AGNES
- Eventually each node forms a cluster on its own



CURE (Clustering Using REpresentatives)

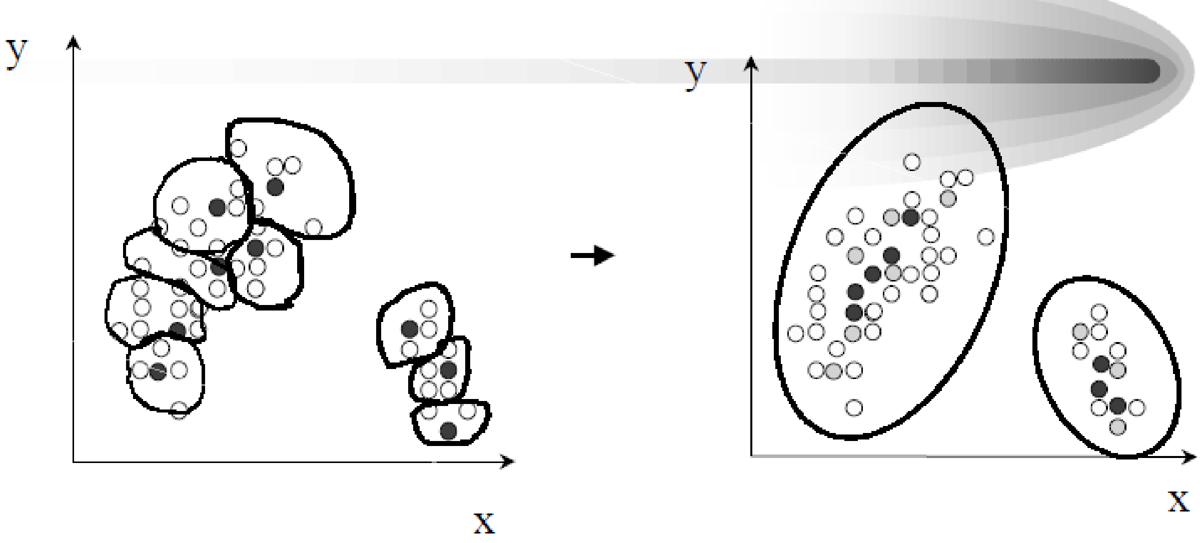


- CURE: proposed by Guha, Rastogi & Shim, 1998
  - Stops the creation of a cluster hierarchy if a level consists of *k* clusters
  - Uses multiple representative points to evaluate the distance between clusters, adjusts well to arbitrary shaped clusters and avoids single-link effect

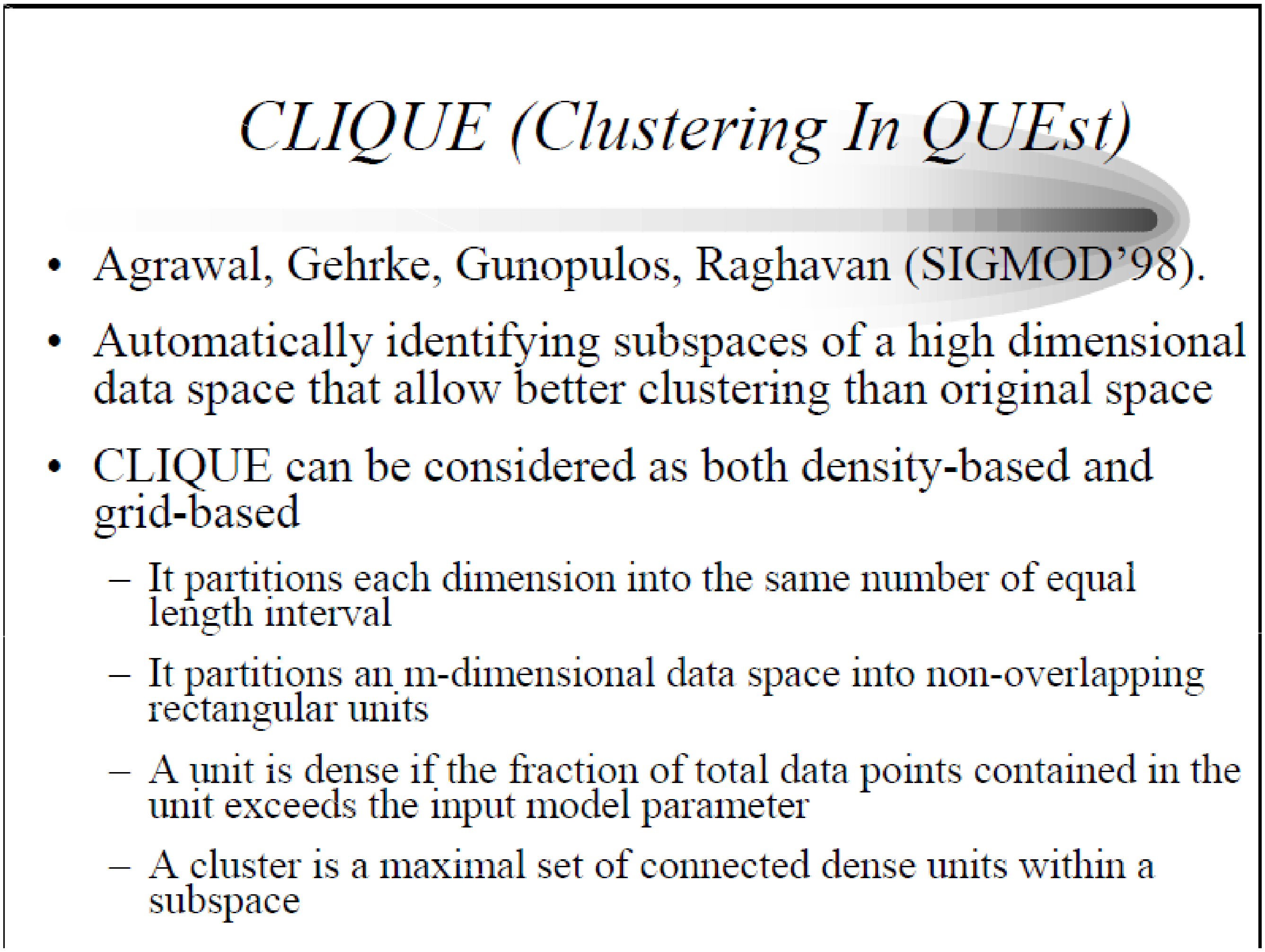
Cure: The Algorithm

- Draw random sample *s*.
- Partition sample to *p* partitions with size *s/p*
- Partially cluster partitions into *s/pq* clusters
- Eliminate outliers
  - By random sampling
  - If a cluster grows too slow, eliminate it.
- Cluster partial clusters.
- Label data in disk

Cure: Shrinking Representative Points



- Shrink the multiple representative points towards the gravity center by a fraction of α.
- Multiple representatives capture the shape of the cluster



CLIQUE (Clustering In QUEst)

- Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98).
- Automatically identifying subspaces of a 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 interval
  - It partitions an m-dimensional data space into non-overlapping rectangular units
  - A unit is dense if the fraction of 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

CLIQUE: The Major Steps

- 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

Strength and Weakness of CLIQUE

- Strength
  - 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
- Weakness
  - The accuracy of the clustering result may be degraded at the expense of simplicity of the method

PROCLUS

- PROCLUS is a top-down approach
- Samples the data and selects a set of k medoids and iteratively improves the subspace clusters.
- 3 phase implementation of algorithm.

- INITIALISATION PHASE:

  Choose sample set of points.

  Choose a set of data point which is probably the medoid.

  (Uses greedy algorithm to do so).

PROCLUS

- ITERATION PHASE:

  looks to find the best medoids from previous phase..

  Replace bad ones with the good medoids.

  For each data point ,assign it to the medoid mi if its manhattan distance is minimum.
- REFINEMENT PHASE:

  Computes new dimensions for each medoid based on clusters and reassigns points to medoids,removing outliers.
- PROCLUS is baiased to spherical data.
- Faster than CLIQUE.

FREQUENT PATTERN BASED CLUSTERING METHODS

- Definition of pattern matrix:

  Submatrix that follows a particular pattern-rows change in a synchronised way w.r.t the columns and vice versa.
- Δ-pcluster-pscore>0
- POPULAR ALGO-**MAPLE**

  Enumerates all maximal pclusters.

  Using Depth first search.

  Apriori pruning of data sets(Monotonicity and apriori)

CLUSTERING STREAMS

- Challenges with streams:

  Massive in size.

  Patterns are continuously changing and evolving.

  Domain based challenges arise-difficult to generalise.

  Offline processing not possible.

CLUSTERING STREAMS

- A simple streaming model:

  Based on k-medians methodology.

  Divide into chunks based on the size of main memory.each chunk is represented as one point in space.

  Sliding window consists of most recent N points.

  BDMO ALGORITHM:

  1.Summarised by buckets whose size is a power of 2.

  2.Only one or two buckets of same size.

  3.Bucket size has to be non-decreasing.

  4.Contents of bucket-size,timestamp,collection of records(cluster points)

CLUSTERING STREAMS

INITIALIZING BUCKETS:

Smallest bucket of size 2(p).

Timestamp is the most recent point’s timestamp in the bucket.

The centroid for the cluster which is added to the record.

MERGING BUCKETS:

Drop if timestamp greater than N time units prior to current time.

If three buckets of same size,merge oldest 2-recursive merging.