A Review on Cluster Based Approach in Data Mining

M. Vijaya Maheswari
PhD Research Scholar,
Department of Computer Science
Karpagam University
Coimbatore, Tamilnadu ,India

Dr T. Christopher
Assistant professor,
Department of Computer Science
Govt. College of Arts & Science
Coimbatore, Tamilnadu ,India

***Abstract*— Data mining refers to extracting or mining knowledge from large databases. Clustering is a data mining technique used for grouping a set of physical objects into classes of similar objects. The choice of clustering algorithm depends on the type of data available and on the particular purpose and application. This paper presents detailed approach on various clustering methods and algorithms.**

Keywords*— *Data Mining; Clustering; Algorithms; Partitioning

INTRODUCTION

Data mining and knowledge discovery in databases is a new interdisciplinary field, merging ideas from statistics, machine learning, database and parallel computing [1, 3]. Data mining is the non-trival extraction of implicit, preciously unknown and potentially useful information from the data.[3]

Clustering is a data mining technique of grouping data into different groups, so that data in each group share similar trends and patterns.[3] Finding similarities between data according to the characteristics found in the data and grouping similar objects into clusters.[8] Clustering is an example of unsupervised learning. It is a form of learning by observation, instead of learning by examples.[1,3]

I. OBJECTIVES OF CLUSTERING

- To uncover natural groupings
- To initiate hypothesis about the data
- To find consistent and valid organization of the data.

Clustering is an emerging field of research where its potential applications pose their own requirements. The following are typical requirements of clustering in data mining. [2]

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Ability to deal with noisy data
- Insensitivity to the order of input records
- High dimensionality
- Constraint based clustering
- Interpretability and usability [7]

II. TYPES OF DATA IN CLUSTER ANALYSIS

A. *Data Matrix (object-by-variable structure)*

This represents n objects, such as persons, with p variables also called measurements or attributes such as height, age, weight, complexion, gender and so on.

B. *Dissimilarity Matrix (object-by-object structure)*

It stores a collection of data's that are available for all pairs of n objects. [1]

- Quantitative data – ex: weight, marks, height, price, salary and count.
- Binary data – ex: gender, marital status
- Qualitative nominal data – which may take more than two values but has no natural order. Ex: religion, foods, colours
- Qualitative ordinal data- similar to nominal data except that the data has an order associated with it. Ex: grades A, B, C, and D, sizes S, M, L, and XL. [7]

III. DISTANCE MEASURES

Many clustering algorithms require that the distance between clusters rather than elements be determined. Given clusters K_i and K_j, there are several standard alternatives to calculate the distance between clusters. [8]

- Single link: The smallest distance between an element from one cluster and an element from the other.
- Complete link: Largest distance between an element in one cluster and an element in the other.
- Average: The average distance between an element from one cluster and an element from the other.
- Centroid: If clusters have a representative centroid, then the distance is defined as the distance between the centroids.
- Medoid: Using a medoid to represent each cluster, the distance between the clusters can be defined by the distance between the medoids.[8]

IV. CLUSTERING METHODS

A. *Partitioning Methods*

- Given a database of n objects or data tuples, partitioning method constructs k partitions of the data, where each partition represents a cluster and K<= n.

- It classifies the data into k groups, which together satisfy the requirements.
- Each group must contain atleast one object.
- Each object must belong to exactly one group.[1]
- K-means algorithms, where each cluster is represented by the center of gravity of the cluster.[3]
- K-medoid algorithms, where each cluster is represented by one of the objects of the cluster located near the center.[3]
- *Types*
  - Agglomerative – merge the closest pair of clusters until only one cluster left.
  - Divisive – Split a cluster until each cluster contains a point**.**
  - Example – CURE, BIRCH, CHAMELEON [1]

B. *Density Based Methods*

- Some clustering methods have been developed based on the notion of density.
- It is to continue growing the given cluster as the density in the neighbourhood exceeds some threshold.
- Such a method can be used to filter out noise and discover clusters of arbitrary shape.
- Example: DBSCAN, OPTICS, DENCLUE.[1]

C. *Grid Based Methods*

- In this class of methods, the object space rather than the data is divided into a grid.
- It is based on characteristics of the data and such methods can deal with non-numeric data more easily.
- Example: STING, CLIQUE, WaveCluster.[2]

D. *Model Based Methods*

- These methods hypothesize a model for each of the clusters and find the best fit of the data to the given model.
- Model based algorithms may locate clusters by constructing a density function that reflects the spatial distribution of the data points.
- It automatically determines the number of clusters based on standard statistics, taking “noise” or outliers into account and thus yielding robust clustering methods.
- Ex: COBWEB, CLASSIT.[1]

V. CLUSTERING ALGORITHMS

A. *Partitioning Algorithms*

1. CENTROID- BASED TECHNIQUE: K-MEANS

It is also called the centroid method since at each step the centroid point of each cluster is assumed to be known and each of the remaining points are allocated to the cluster whose centroid is closest to it. [7]

Steps for Implementation

- Partition objects into k nonempty subsets.
- Then Compute the seed points as the centroids for the clusters of the current partition.
- Then assign each object to the cluster with the nearest seed point.
- Then go back to step 2, else stop when no more new assignment.[1]

2. REPRESENTATIVE OBJECT BASED TECHNIQUE: K- MEDOIDS

This algorithm is very similar to k-Means with the small exception of instead of creating an artificial point to recalculate the mean point, k-Medoids recalculates from the nearest actual point in a data set.[1]

Steps for Implementation

- Arbitrarily choose k objects in D as the initial representative objects or seeds.
- Repeat
- Assign each remaining object to the cluster with the nearest representative object
- Randomly select a non-representative object, $o_{random.}$
- Compute the total cost, S, of swapping representative object o_p with $o_{random.}$
- If S < 0 then swap o_j with o_{random} to form the new set of k representative objects
- Until no change.[1]

3. FROM K-MEDOIDS TO CLARANS

In case of CLARA, instead of taking the whole set of data into consideration, only a small portion of the real data is chosen as a representative of the data, and medoids are chosen from this sample using PAM. It then classifies the remaining objects using the partitioning principle.

Steps for Implementation

- Randomly choose k medoids
- Randomly consider one of the medoids to be swapped with a non- medoid
- If the cost of the new configuration is lower, repeat step 2 with new solution
- If the cost is higher, repeat step 2 with different non-medoid object, unless a limit has been reached
- Compare the solutions so far, and keep the best
- Return to step 1, unless a limit has been reached [4]

B. *Hierarchical Algorithms*

1. BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)

This algorithm first partitions objects hierarchically using tree structures and then applies other clustering algorithms to form refined clusters.

Phases

- BIRCH scans the database to build an initial in- memory CF tree, which can be viewed as multilevel compression of the data. It tries to preserve the inherent clustering structure of the data.

- BIRCH applies the selected clustering algorithm to cluster the leaf nodes of the CF tree, which removes sparse clusters as outliers and groups dense clusters into larger ones.[1]

2. *CURE (Clustering Using Representatives)*

CURE measures the similarity of both clusters based on the similarity of the closest pair of the representative points belonging to different clusters, without considering the internal closeness of the two clusters involved. Creates clusters by sampling the database and shrinks them toward the center of the cluster by a specified fraction. [2]

Steps

- Draw a random sample S, of the original objects.
- Partition sample S into a set of partitions and form a cluster for each partition.
- Representative points are found by selecting a constant number of points from a cluster and then "shrinking" them toward the center of the cluster.
- Cluster similarity is the similarity of the closest pair of representative points from different clusters.
- Shrinking representative points toward the center helps avoid problems with noise and outliers
- CURE is better able to handle clusters of arbitrary shapes and sizes [1]

3. *CHAMELEON (A Hierarchical Clustering Algorithm Using Dynamic Modelling)*

The merging process used by the CHAMELEON using the dynamic model and facilitates discovery of natural homogenous clusters. It operates on a sparse graph in which nodes represent data items, and weighted edges represent similarities among the data items. [2, 5]

Steps

- Pre-processing- represent the data by a graph
- Given a set of points, construct the k-nearest neighbour (k-NN) graph to capture the relationship between a point and its k nearest neighbours.
- Concept of neighbourhood is captured dynamically.
- *PHASE 1*: This phase use a multilevel graph partitioning algorithm on the graph to find a large number of clusters of well-connected vertices.
- *PHASE 2*: This phase use hierarchical agglomerative clustering to merge sub-clusters.[1]

C. Density Based Algorithms

1. DBSCAN(Density Based Spatial Clustering of Applications with Noise)

Cluster is defined as a maximal set of density connected points. It discovers clusters of arbitrary shape in spatial databases with noise.

Algorithm

- Arbitrary select a point p
- Retrieve all points density- reachable from p.
- If p is a core point, a cluster is formed.
- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

2. OPTICS (Ordering Points To Identify the Clustering Structure)

It produces a special order of the database with its density based clustering structure. It can be represented graphically using visualization techniques.[1, 5]

D. Grid Based Algorithms

1. STING (Statistical Information Grid Approach)

The spatial area is divided into rectangular cells. There are several levels of cells corresponding to different levels of resolution.

Method

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level.
- Statistical info of each cell is calculated and stored beforehand and is used to answer queries.
- Parameters of higher level cells can be easily calculated from parameters of lower level cell.
- Use a top down approach to answer spatial data queries
- Start from pre-selected layer
- For each cell in the current level compute the confidence interval. [1, 4]

2. *WAVECLUSTER*

A signal processing technique that decomposes a signal into different frequency sub-band. Data are transformed to preserve relative distance between objects at different levels of resolution. It allows natural clusters to become more distinguishable.[1, 6]

E. Model Based Algorithms

1. COBWEB

It is a popular method of incremental conceptual learning. This method creates a hierarchical clustering in the form of a classification tree. Each & every node refers to a concept and contains a probabilistic description of that concept.

Limitations

- The assumption is 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's.

2. *CLASSIT*

It is an extension of COBWEB for incremental clustering of continuous data. It suffers similar problems as COBWEB. [1]

CONCLUSION

This paper presents a detailed study about clustering method in data mining. It also explained about various clustering algorithms such as partitional, hierarichal, density based, grid based and model based algorithms. The algorithms represent is limited to steps but not with explanation by working with example.

REFERENCES

[1] B.S Charualtha, S. Poonkuzhali, C.Saravanakumar, "DATA WAREHOUSING AND DATA MINING", Charulatha Publications, 2014.

[2] A.K Jain and R.C Dubes, "Algorithms for Clustering Data", Prentice Hall, 1998.

[3] Arun K Pujari , "Data Mining Techniques", Universities press, 2001.

[4] Jaiwei Han , Micheline Kamber, "Data Mining concepts and techniques", Second Edition, Morgan Kaufmann series,2005.

[5] K.P.Soman, Shyam Diwakar , V.Ajay, "Insight into Data Mining Theory and Practice", PHI Publications, 2006.

[6] Richard J.Roiger, Michel W.Geatz, " Data Mining- A Tutorial – Based Primer", Pearson Education Inc., 2003.

[7] G.K.Gupta, "Introduction to Data Mining with Case Studies", Prentice-Hall of India, 2008.

[8] Margaret H.Dunham, " Data Mining", Pearson, 2013.