International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869, Volume-3, Issue-4, April 2015

Data Clustering and Algorithm

Seema Yadav

***Abstract*— Clustering is a criteria by which we can group various entities. The entities can be physical entities like a worker and can be a abstract such as the behavior of the worker. Clustering is very useful task into the data mining technology. The main task of data mining is to collect the data into the organized form from the unorganized data set. But there is a major difference between cluster and data mining technology and that is cluster is unsupervised learning and data mining is a supervised learning. So because of this cluster does not require training persons to work on them where data mining technology require experienced persons. Data mining is an area where Computer science, Biology, Artificial Intelligence and Statistics meet and goal to search and retrieve useful information that hidden in data. In data mining, mining of data can be done using two learning approach- Supervised and Unsupervised learning. Clustering is the task of grouping a set of objects in such a way that objects in the cluster are more similar to each other than to those in other clusters based on the principle of maximum interclass similarity and minimum interclass similarity. Clustering faces many challenges such as discovery of clusters with arbitrary shape, dealing with noisy data, high dimensionality, Ability to deal with all types of attributes, requirements for domain knowledge to determine input parameter.**

This paper only describes the clustering and clustering algorithm like K-mean algorithm, Hierarchical Clustering, Density Based clutestering and Expectation Maximization Clustering. This paper only describes the overview of the various types of clustering.

***Index Terms*— K mean clustering algorithm, Hierarchical Clustering, Divisive Hierarchical, Density Based algorithm.**

I. INTRODUCTION

Clustering is the process of grouping of various types of entities. These entities can be physical entities like a worker and can be an abstract like the behavior of the worker. It is the task of collecting a se of objects in the groups are linked or in relation with other cluster. The cluster is not specific algorithm and but solve a particular task. The clustering challenges the data mining in many ways like scalability, ability to deal with different types of attribute, ability to deal with different type of noisy data.

II. CLASSIFICATION OF DIFFERENT TYPES OF CLUSTERING ALGORITHM:-

- K-Mean clustering algorithm:- K-Mean clustering algorithm just like the simple partition algorithm which partition the cluster of the data from a collection of N objects Where K<=N.K-Mean clustering identify the mean value of data point within the cluster.

Algorithm-

1. *Begin: Randomly Choose K data objects from data set D as initial centres.*
   Number of cluster=k:
2. *Repeat*
   a. Assume each cluster as centriod
   b.Calculate the distance of all data point to centriod.
   c.Assign data objects d_i to the nearest cluster
3. *Update for each cluster (1<=j<=k)*
4. *Recalculate the cluster center*
5. *Untill no changes in the center of the cluster.*
6. *End*

- Hierarchical Clustering:- Hierarchical clustering method merged or splits the similar data objects by making hierarchy of cluster also known as Dendogram.Hierarchical clustering is classified into two forms

(1).Agglomerative:-

This clustering is done in bottom up approach which take a single object as a single cluster. The algorithm is

1. *Begin*
   Assign number of cluster= Number of objects
2. *Repeat When number of cluster =1*
   a) Find the minimum intercluster distance
   b) Merge the minimum intercluster distance
3. *End*

(2).Divisive Hierarchical:- It is the top down approach. This clustering start with one cluster that contains all data objects. Then in each successive iteration it divide into the cluster by satisfying some similarity criteria until each objects forms cluster or satisfies stopping cluster.

Algorithm

1. *Assign no of cluster=Number of objects*
2. *Repeat when number of cluster =1*
   a) Find the minimum intercluster distance
   b) Merge the minimum intercluster distance
3. *End*

Density based algorithm:- A cluster is a dense region of points that is separated by low density region. Density based clustering algorithm is used when the cluster are irregular. It identifies core objects and connects the core objects and their neighborhood to form dense region as cluster.

Algorithm:-

1. Select an arbitrary point r
2. Retrieve the neighborhood of r using '€'/
3. If the density of neighbor reach to the threshold, clustering process start else points is mark as a noise.
4. Repeat the process until all of the points have been processed.

Manuscript received April 15, 2015.
Seema Yadav, WCTM student farrukhnagar, Gurgaon, Haryana

ACKNOWLEDGEMENT

I would like to thank my supervisor, Ms.Pragati, for providing me with supervision, motivation and continuous encouragement throughout the course of this work. He gives me to the correct direction at every stage of the research Without his supervision and simulating guidance, I would not be able to complete this work.

REFERENCES

[1] By Amandeep Kaur Mann & Navneet Kaur
[2] A.K Jain (Michigan State University)
[3] M.N Murty (Indian Institute of Science)
[4] Prof.Neha Soni (Gujraat Technological University)
[5] P.J Flynn (The Ohio state university)