IJCSE International Journal of Computer Science and Engineering Open Access
Review Paper Volume-3, Issue-7 E-ISSN: 2347-2693

Clustering Algorithms in Data Mining: A Comprehensive Study

Jaskaranjit Kaur and Gurpreet Kaur*

Dept. of CSE, DAV University, Jalandhar,India

www.ijcseonline.org

Received: Jun/16/2015 Revised: Jun/28/2015 Accepted: July/19/2015 Published: July/30/ 2015

***Abstract*—** Distribution of dataset into a set of homogeneous clusters is the elementary operation in data mining. Clustering is the key technique of distribution in data mining. Clustering is the method of grouping data objects in such a way that the data objects in the same cluster are more similar (intra-cluster similarity) to each other and are less similar to data objects in other cluster (inter-cluster similarity). Clustering can be done with different algorithms. In this review paper, a survey of clustering and its different techniques is done. This paper covers some of the partitioning and hierarchical clustering algorithms.

***Keywords*—** Clustering, Clustering Techniques, Partitioning Clustering, Hierarchical Clustering, K-Means, K-Medoid, Birch

I. INTRODUCTION

Data mining is a way of analyzing datasets for the purpose of finding out unsuspected relationships and to summarize the data in better ways. The information obtained should be clear, understandable and useful to us [3]. It is a method of extracting information from large volumes of raw data. Clustering is one of the main task in data mining applications. Clustering is an unsupervised learning technique where one attempts to identify a finite set of categories which are called clusters to describe the data in better way [2]. The clusters found after applying clustering algorithms should have high intra-cluster similarity and low inter-cluster similarity. Intra-cluster similarity means similarity to objects within the same cluster and inter-cluster similarity is dissimilarity to objects in other clusters [1]. This review paper focuses on clustering in data mining. Data mining attempts to cluster the complications of very large datasets having many attributes of different types. This leads to the requirements of clustering algorithms. A large variety of clustering algorithms emerged recently that meet these requirements. These algorithms are subject of the survey in this paper.

In data mining two learning approaches are used to mine data which are supervised learning and unsupervised learning.

Supervised Learning: Supervised learning is also known as directed data mining because in this the training data includes both the input and the desired results. In case of some examples the correct result are known and are given in input to model during the learning process. These techniques are usually accurate and fast [1].

Unsupervised Learning: Unsupervised learning is also known as undirected data mining. In this the model is not provided with the correct results duri ng the training. This technique can be used to cluster the input data in classes on the basis of some statistical properties of the data. The labeling is done after cluster formation by the clustering algorithms [1].

Clustering is an unsupervised learning technique that partition data objects into a number of groups, such that data objects in the same cluster are more similar to each other and data objects in different clusters are dissimilar, according to some criteria. Different from supervised learning, where training examples are associated with a class label that expresses the membership of every example to a class, clustering assumes no information about the distribution of the objects and it has the task to both discover the classes present in the data set and to assign objects among such classes in the best way [2].

This paper includes two types of clustering algorithms 1) Partitioning clustering algorithms 2) hierarchical clustering algorithms.

II. CLUSTERING METHODS

The Clustering Methods described in this paper can be classified into following main categories [6].

- Partitioning Clustering
- Hierarchical Clustering

A. Partitioning Clustering

Suppose we are having a dataset of n objects, the partitioning algorithm divides data into k partitions where k<n. Each partition is a cluster .It means that it will classify the data into k groups, which satisfy the following requirements [1]:

- Each group contains at least one object.
- Each object must belong to exactly one group.

❖ For a given number of partitions (say k), the partitioning method will create an initial partitioning.

- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other [2].

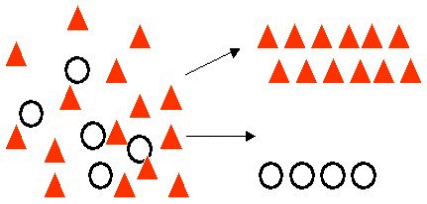

Fig1: Partitioning Clustering

Advantages of Partitioning clustering [2]
1. Easy to implement.
2. Suitable for datasets with compact spherical clusters that is well-separated

Disadvantages [2]
1. Need to specify K, the number of clusters
2. Local minimum Initialization matters
3. Empty clusters may appear

B. *Hierarchical Clustering*

This method creates the hierarchical decomposition of the given set of data objects. We can classify Hierarchical method on basis of how the hierarchical decomposition is formed as follows [7,11]:

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach: This approach is also known as bottom-up approach. In this we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds [7].

Divisive Approach: This approach is also known as top-down approach. In this we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds [7].

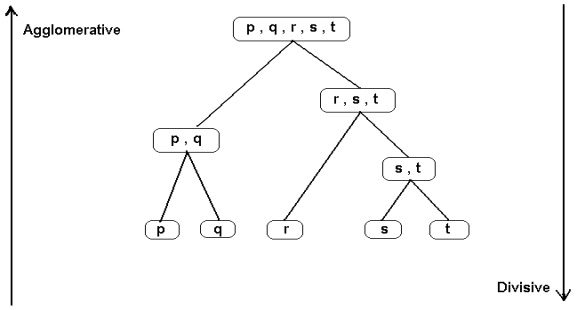

Fig2: Hierarchical clustering

Advantages of Hierarchical Clustering [2]
1) No apriori information about the number of clusters required.
2) Easy to implement and gives best result in some cases.

Disadvantages [2,7]
1) Algorithm can never undo what was done previously.
2) Time complexity of at least $O(n^2 \log n)$ is required, where *'n'* is the number of data points.
3) Based on the type of distance matrix chosen for merging different algorithms can suffer with one or more of the following:
   i) Sensitivity to noise and outliers
   ii) Breaking large clusters
   iii) Difficulty handling different sized clusters and convex shapes
4) No objective function is directly minimized
5) Sometimes it is difficult to identify the correct number of clusters by the dendogram.

C. *Hierarchical vs Partitioning Clustering*

Clustering is a machine learning technique for analyzing data and dividing into groups of similar data. These groups or sets of similar data are known as clusters. Cluster analysis looks at clustering algorithms that can identify clusters automatically. Hierarchical and Partitional are two such classes of clustering algorithms. Hierarchical clustering algorithms break up the data in to a hierarchy of clusters. Paritional algorithms divide the data set into mutually disjoint partitions [7]. Hierarchical and Partitional Clustering have key differences in running time, assumptions, input parameters and resultant clusters. Typically, partitional clustering is faster than hierarchical clustering. Hierarchical clustering requires only a similarity measure, while partitional clustering requires stronger assumptions such as number of clusters and the initial centers. Hierarchical clustering does not require any input parameters, while partitional clustering algorithms require the number of clusters to start running. Hierarchical clustering returns a much more meaningful and subjective division of clusters but partitional clustering results in exactly k clusters. Hierarchical clustering algorithms are more suitable for categorical data as long as a similarity measure can be defined accordingly [11].

III. PARTITIONING CLUSTERING ALGORITHMS

This section describes the K-Means and k-Medoid clustering algorithms

A. *K-Means Clustering*

K-means (MacQueen'67): Each cluster is represented by the center of the cluster. It is an algorithm to group or to classify our objects based on attributes/features into K number of group. K is positive integer number. The grouping is done based on minimizing the sum of squares of distances between data and the corresponding cluster centroid. K-Means is a numerical, iterative method. It is one of the simplest unsupervised learning algorithms. The main concept of the algorithm is to define k centroids, one for each cluster [5]. These centroids should be positioned in a scheming way since different location of k causes different cluster result. Then the K-Means algorithm will iteratively cluster data to find the k centroids and assign each object to the nearest centroid where the centroid is the mean of the coordinates of the objects in the cluster. Then, the k centroids will change their positions step by step until no further changes occur. The k-means algorithm is as follows [4]:

1. Select k points as initial centroids (randomly generated vectors can also be used).
2. Calculate the distance from each cluster centroid to each point.
3. Assign each point to the nearest cluster.
4. Calculate new cluster centroid, where each new centroid is the mean of all vectors in that cluster.
5. Repeat steps 2-4 until a stopping condition is reached.

B. K-Medoid clustering algorithm

k-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster .The k-Means algorithm is sensitive to outliers as an object with an enormously large value may significantly twist the distribution of data [3]. In K-Medoid clustering instead of taking the mean value of the data objects in a cluster as a reference point, a medoid can be used, which is the most centrally located object in a cluster. Thus, the partitioning method can still be performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point. This is the basic concept of the k-Medoids algorithm. The basic strategy of k- Medoids clustering algorithms is to find k clusters by first at random finding a representative object (the medoids) for each cluster. Each remaining object is clustered with the medoid to which it is the most similar. The k-Medoids method uses representative objects as reference points instead of taking the mean value of the objects. The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects.

K-Mediods algorithm for partitioning based on medoid is as follows [3]:

1. Arbitrarily choose k objects in D as the initial representative objects.
2. **Repeat** assign each remaining object to the cluster with the nearest medoid.
3. Randomly select a non medoid object Orandom.
4. Compute the total points S of swapping object Oj with Oramdom.
5. If S < 0 then swap Oj with Orandom to form the new set of k medoid.

   Until no change.

It attempts to determine k partitions for n objects. After an initial random selection of k medoids, the algorithm repeatedly tries to make a better choice of medoids. Therefore, the algorithm is often called as representative object based algorithm.

IV. HIERARCHICAL CLUSTERING ALGORITHM

This section describes the BIRCH and CURE clustering algorithms.

A. BIRCH

Balanced Iterative Reducing and Clustering Using Hierarchies. It is an agglomerative Clustering technique designed for clustering a large amount of numerical data [11].

What Birch algorithm tries to solve?

- Most of the existing algorithms do not consider the case that datasets can be too large to fit in main memory.
- They do not concentrate on minimizing the number of scans of the dataset
- I/O costs are very high

Key components used in Birch algorithm are [6]:

- *Clustering Feature (CF)***:** It contains summary of the statistics for a given cluster, the 0-th, 1st and 2nd moments of the cluster from the statistical point of view. It is used to compute centroids, and measures the compactness and distance of clusters
- *CF-Tree:* It is also known as height-balanced tree. Two parameters are used in CF-Tree first is the number of entries in each node. Second is the diameter of all entries in a leaf node.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm [9] is an integrated hierarchical clustering algorithm. It uses the clustering features (Clustering Feature, CF) and cluster feature tree (CF Tree) two concepts for the general cluster description. Clustering feature tree outlines the clustering of useful information, and space is much smaller than the meta-data collection can be stored in memory, which can improve the algorithm in clustering large data sets on the speed and scalability and is very suitable for handling discrete and continuous attribute data clustering problem. In the BIRCH tree a node is called a Clustering Feature. It is a small representation of an

underlying cluster of one or many points. BIRCH builds on the idea that points that are close enough should always be considered as a group [6].
Clustering Features are stored as a vector of three values: CF = (N; LS; SS). The linear sum (LS), the square sum (SS), and the number of points it encloses (N) [6].

$$\overrightarrow{LS} = \sum_{i=1}^{N} \overrightarrow{X_i} \quad (1)$$

$$SS = \sum_{i=1}^{N} \overrightarrow{X_i^2} \quad (2)$$

If divided by the number of points in the cluster the linear sum marks the centroid of the cluster. As the formulas suggest that both of these values can be computed iteratively. Any Clustering Feature in the tree can be calculated by adding its child Clustering Features [6]:

$$CF_1 + CF_2 = (\boldsymbol{N_1 + N_2}, \overrightarrow{\boldsymbol{LS_1}} + \overrightarrow{\boldsymbol{LS_2}}, \boldsymbol{SS_1 + SS_2}) \quad (3)$$

A CF tree is a height balanced tree that has two parameters namely, a branching factor, B, and threshold, T.
B = Branching Factor, maximum children in a non-leaf node
T = Threshold for diameter or radius of the cluster in a leaf
The Birch clustering algorithm works in four phases [6].

In phase1, the initial CF is built from the database based on the branching factor B and the threshold value T.
Phase2 is an optional phase in which the initial CF tree would be reduced in size to obtain a smaller CF tree.
Global clustering of the data points is performed in phase3 from either the initial CF tree or the smaller tree of phase2.
As has been shown in the evaluation good clusters can be obtained from phase3 of the algorithm. If it is required to improve the quality of the clusters, phase4 of the algorithm would be needed in the clustering process.

B. CURE

CURE (Clustering Using Representatives)CURE is an agglomerative hierarchical clustering algorithm that creates a balance between centroid and all point approaches. Basically CURE is a hierarchical clustering algorithm that uses partitioning of dataset. A combination of random sampling and partitioning is used here so that large database can be handled. In this process a random sample drawn from the dataset is first partitioned and then each partition is partially clustered. The partial clusters are then again clustered in a second pass to yield the desired clusters. It is confirmed by the experiments that the quality of clusters produced by CURE is much better than those found by other existing algorithms [11].

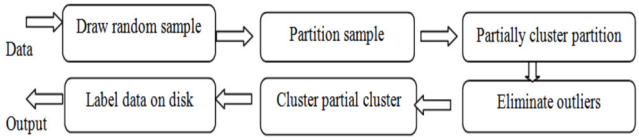

Fig3. Process of CURE Clustering Algorithm

CURE is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. CURE achieves this by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the cluster and then shrinking them toward the centre of the cluster by a specified fraction [11].

V. CONCLUSION

The key objective of data mining is to get the information from a large data set and convert it into a form which is understandable and useful in future. Clustering is an important technique in data mining which groups a set of data objects so that the data objects in same cluster have high intra-cluster similarity with each other and low inter-cluster similarity with the data objects of other clusters. There are various algorithms in data mining to cluster the data. In this review paper two types of clustering techniques are discussed in detail. So this paper provides a quick review of partitioning and hierarchical clustering algorithm. In future these two clustering techniques can be combined or merged to propose a new clustering algorithm.

VI. REFERENCES

[1]. K.Kameshwaran, K.Malarvizhi, "Survey on Clustering Techniques in Data Mining", 2014 International Journal of Computer Science and Information Technologies, Vol. 5 (2) , **2014**, 2272-2276
[2]. Deepti Sisodia, Lokesh Singh, Sheetal Sisodia, Khushboo saxena, "Clustering Techniques: A Brief Survey of Different Clustering Algorithms" **2012** International Journal of Latest Trends in Engineering and Technology, Vol. 1, Issue 3 ,September 2012
[3]. Saurabh Shah,Manmohan Singh, "Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm", **2012** IEEE International Conference on Communication Systems and Network Technologies.
[4]. Shalove Agarwal, Shashank Yadav, Kanchan Singh, "K-means versus K-means ++ Clustering Technique", **2012** IEEE Second International Workshop on Education Technology and Computer Science

[5]. Shi Na , Liu Xumin, Guan yong , "Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm", **2010** IEEE Third International Symposium on Intelligent Information Technology and Security Informatics.
[6]. NidalIsmael, Mahmoud Alzaalan , WesamAshour, "Improved Multi Threshold Birch Clustering Algorithm" 2014 International Journal of Artificial Intelligence and Applications for Smart Devices, Vol.2 , No.1 (**2014**), pp.1-10.
[7]. J. Han and M. Kamber,"Data Mining: concepts and techniques", Beijing: China Machine Press, Third Edition (**2012**).
[8]. R.Xu, "Survey of Clustering Algorithms" **2005** IEEE Trans.Neural Networks.
[9]. Archana Singh, Avantika Yadav, Ajay Rana, "K-means with Three different Distance Metrics", **2013** International journal of computer Applications.
[10]. S. N. Alsaleh, R. , Yue Xu, "Grouping people in social networks using a weighted multi-constraints clustering method ", 2012 IEEE InternationalConference on Fuzzy Systems (FUZZ-IEEE), pp. 1-8, **2012**
[11]. Yogita Rani, Manju , Harish Rohil, " Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9" , 2014 The SIJ Transactions on Computer Science Engineering & its Applications, (CSEA) Volume 3, Number 10 (**2013**), pp. 1115-1122.
[12]. RichaDhiman, ShevetaVashisht, "A Cluster analysis and Decision Tree Hybrid Approach in Data Mining to Describe Tax Audit", International Journal of Computers & Technology Volume 4 No. 1, Jan-Feb, **2013**
[13]. Manpreet Kaur, Usvir Kaur, " Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection" 2013 International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July **2013**

