Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 8, August 2017, pg.109 – 115

A Comparative Study of Various Clustering Algorithms in Data Mining

K. Chitra[1], Dr. D.Maheswari[2]

Research Scholar, School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, Sulur, Coimbatore[1]

Head, Research Coordinator, School of Computer Studies- PG, Rathnavel Subramaniam College of Arts and Science, Sulur, Coimbatore[2]

E mail: chitra.k@rvsgroup.com [1], maheswari@rvsgroup.com [2]

Abstract --The purpose of the data mining technique is to mine information from a bulky data set and make it into a reasonable form for supplementary purpose. Data mining can do by passing through various phases. Mining can be done by using supervised and unsupervised learning. Clustering is a significant task in data analysis and data mining applications. It is the task of arranging a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). The clustering is unsupervised learning. Clustering algorithms can be classified into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. This paper focuses on a keen study of different clustering algorithms in data mining. A brief overview of various clustering algorithms is discussed.

Keywords: data mining, clustering, clustering algorithms, techniques

I. INTRODUCTION

Data mining refers to extracting information from large amounts of data, and transforming that information into an understandable and meaningful structure for further use. Data mining is an essential step in the process of knowledge discovery from data (or KDD). It helps to extract patterns and make hypothesis from the raw data. Tasks in data mining include anomaly detection, association rule learning, classification, regression, summarization and clustering [1].

In Data Mining the two types of learning sets are used, they are supervised learning and unsupervised learning.

a) Supervised Learning

In supervised training, data includes together the input and the preferred results. It is the rapid and perfect technique. The accurate results are recognized and are given in inputs to the model through the learning procedure. Supervised models are neural network, Multilayer Perceptron and Decision trees.

b) Unsupervised Learning

The unsupervised model is not provided with the accurate results during the training. This can be used to cluster the input information in classes on the basis of their statistical properties only. Unsupervised models are for dissimilar types of clustering, distances and normalization, k-means, self organizing maps.

II. CLUSTERING

Clustering is a major task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups. Cluster is an ordered list of data which have the familiar characteristics. Cluster analysis can be done by finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. Clustering is an unsupervised learning process. Clustering has an extensive and prosperous record in a range of scientific fields in the vein of image segmentation, information retrieval and web data mining.

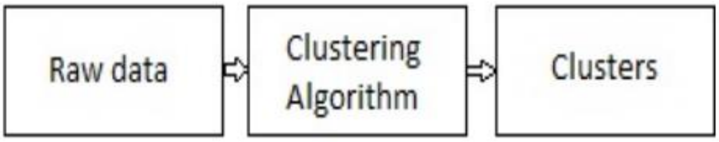


Figure 1: Stages of Clustering [15]

III. CLUSTERING ALGORITHMS

Clustering algorithms can be categorized into partition-based algorithms hierarchical-based algorithms, density-based algorithms and grid-based algorithms. These methods vary in (i) the procedures used for measuring the similarity (within and between clusters) (ii) the use of thresholds in constructing clusters (iii) the manner of clustering, that is, whether they allow objects to belong to strictly to one cluster or can belong to more clusters in different degrees and the structure of the algorithm. Irrespective of the method used, the resulting cluster structure is used as a result in itself, for inspection by a user, or to support retrieval of objects [5].

A. Partitioning Algorithms

Partitioning clustering algorithm split the data points into k division, where each division represent a cluster and k<=n, where n is the number of data points. Partitioning methods are based on the idea that a cluster can be represented by a centre point. The cluster must exhibit two properties, they are (a) each collection should have at least one object (b) every object should belong to accurately one collection. The main drawback of this algorithm is whenever a point is close to the center of another cluster; it gives poor outcome due to overlapping of data points [4]. It uses a number of greedy heuristics schemes of iterative optimization.

There are many methods of partitioning clustering; they are K-Means, K-Medoids Method, PAM (Partitioning around Medoids), and CLARA (Clustering Large Applications)[8].

1) K-Means algorithm: In this algorithm a cluster is represented by its centroid, which is a mean (average pt.) of points within a cluster [3]. This works efficiently only with numerical attributes. And it can be negatively affected by a single outlier. The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications. The technique aims to partition “n” observations into *k* clusters in which every observation belongs to the cluster with the nearby mean.

The basic algorithm is very simple

1. Select K points as initial centroids.
2. Repeat.
3. Form K clusters by assigning each point to its closest centroid.
4. Re-compute the centroid of each cluster until centroid does not change.

The *k*-means algorithm has the following significant properties:

1. It is effective in dealing out huge data sets.
2. It frequently terminates at a local optimum.
3. It works just on numeric values.
4. The clusters have convex shapes [11].

Disadvantages of k-means

1. Generally terminates at the local optimum, and not the global optimum.
2. Can only be used when the mean is defined and therefore requires specifying k, the number of clusters, in advance.

2) K-Medoids Algorithm: In this algorithm we utilize the actual entity to represent the cluster, using one representative entity per cluster. Clusters are generated by points which are close to respective methods. The partitioning is made based on minimizing the sum of the dissimilarities among every object and its cluster representative [13].

PAM: Like all partitioning methods, PAM works in an iterative, greedy way. The initial representative objects are chosen randomly, and it is considered whether replacing the representative objects by non-representative objects would improve the quality of clustering. The representative objects are replaced with other objects and it continues until the quality cannot be enhanced further. PAM searches for the best k-medoids among a given data set. The time complexity of PAM is O(k(n-k)2) [13]. For large values of n and k, this computation becomes even more costly than the k-means method.
Algorithm [13]:

```
a)  Randomly choose k objects in D as the first representative objects or seeds.
b)  Repeat
    i) Allocate each lasting object to the cluster with the nearby representative object
    ii) Arbitrarily choose a non-representative object, $o_{random}$
    iii) Calculate the total cost, S, of swapping representative objects oj with $o_{random}$
    iv) If S<0 then swap oj with $o_{random}$ to form the new set of k representative objects.
c) Until no change
```

CLARA: To deal with huge data sets, a sampling based technique, called CLARA (Clustering Large Applications) and an improved version which is based on randomized search called CLARANS (Clustering Large Applications based upon Randomized Search) can be used[14].

CLARA uses 5 samples, each with 40+2k points, each of which are then subjected to PAM, which computes the best medoids from the sample [2]. A large sample usually works well when all the objects have equal probability of getting selected. The complexity of computing medoids from a random sample is O(ks2+k(n-k)), Where s is the size of the sample [13].

CLARA cannot find a good clustering if any of the best sampled medoids is far from the best k-medoids.

Pros and Cons of Partitioning Algorithm: It is simple to understand and implement. It takes less time to execute as compared to other techniques. The drawback of this algorithm is the user has to provide pre-determined value of k and it produces spherical shaped clusters. It cannot handle with noisy data objects

B. Hierarchical Algorithm

This algorithm partitions the related dataset by constructing a hierarchy of clusters. It uses the distance matrix criteria for clustering the data. It constructs clusters step by step. In hierarchical clustering, in single step, the data are not partitioned into a particular cluster. It takes a sequence of partitions, which may run from a single cluster containing all objects to ‘n’ clusters each containing a single object. Hierarchical Clustering is classified as

1. Agglomerative Nesting
2. Divisive Analysis

1) Agglomerative Nesting: It is also known as AGNES. It is bottom-up approach. This method construct the tree of clusters i.e. nodes. The criteria used in this method for clustering the data is min distance, max distance, avg distance, center distance. The steps of this method are:

(1) Initially all the objects are clusters i.e. leaf.

(2) It recursively merges the nodes (clusters) that have the maximum similarity between them.

(3) At the end of the process all the nodes belong to the same cluster i.e. known as the root of the tree structure.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies): BIRCH is an agglomerative hierarchical based clustering algorithm. It is used for clustering large amounts of data. It is based on the notion of a clustering feature (CF) and a CF tree. A CF tree is a height-balanced tree. Leaf nodes consist of a sequence of clustering features, where each clustering feature represents points that have already been scanned. It is mainly used when a small number of I/O operations are needed. BIRCH uses a multi clustering technique, wherein a basic and good clustering is produced as a result of the first scan, and additional scans can be used to further improve the quality of clustering [13]. The time complexity of BIRCH is O(n) where n is number of clusters [13].

CHAMELEON: CHAMELEON is an agglomerative hierarchical clustering technique where, unlike other algorithms which use static model of clustering, CHAMELEON uses dynamic modeling to merge the clusters. It does not need users to provide information, but adjusts the clusters to be merged automatically according to their intrinsic properties. It has a complexity of O(Nm+Nlog(N)+m2log(m)), where N is the number of data points and m is the number of partitions [16]. CHAMELEON uses the interconnectivity and compactness of the clusters to find out the similarity between them. It can be used to find clusters of varying densities [12]. However, the processing time for high-dimensional data may go up to O(n2).

2) Devise Analysis: It is also known as DIANA. It is top-down approach. It is introduced in Kaufmann and Rousseeuw (1990). It is the inverse of the agglomerative method. Starting from the root node (cluster) step by step each node forms the cluster (leaf) on its own. It is implemented in statistical analysis packages, e.g., plus.

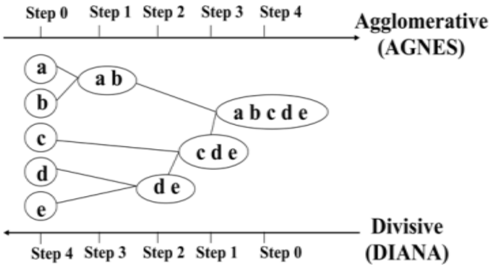


Figure 2: Representation of AGNES and DIANA

Pros and Cons of hierarchical clustering [2]

It has embedded flexibility with regard to the level of granularity and it is easy to handle any forms of similarity or distance and applicable to any attributes type. The drawback of this algorithm is most hierarchical algorithm does not revisit once constructed clusters with the purpose of improvement.

C. Density Based Algorithms

Density based algorithms locate the cluster according to the regions which grow with high density. It is the one-scan algorithms. The first approach called the density-based connectivity clustering pins density to a training data point. It is able to find the arbitrary shaped clusters and handle noise. Representative algorithms include DBSCAN, GDBSCAN, OPTICS (Ordering Points to Identify the Clustering Structure), and DBCLASD. The second approach pins density to a point in the attribute space and is called Density Functions. It includes the algorithm DENCLUE.

1) DBSCAN (Density Based Spatial Clustering of Applications with Noise): The density based algorithm DBSCAN is commonly known. The Eps and the Minpts are the two parameters of the DBSCAN [6]. The basic idea of DBSCAN algorithm is that a neighborhood around a point of a given radius (ε) must contain at least minimum number of points (MinPts) [6]. The steps of this method are:

(1) Randomly select a point t

(2) Recover all density-reachable points from t wrt Eps and MinPts.

(3) Cluster is created, if t is a core point

(4) DBSCAN visits the next point of the database. If t is a border point, and no points are density-reachable from t.

(5) Continue the procedure until all of the points have been processed.

2) DENCLUE (Density-based Clustering): Denclue is a clustering method that depends upon density distribution function. DENCLUE uses a gradient hill-climbing technique for finding local maxima of density functions [10]. These local maxima are called density attractors, and only those local maxima whose kernel density approximation is greater than the noise threshold are considered in the cluster. [13]

Pros and Cons of Density-Based Algorithm: The advantage of this algorithm is it does not require a-priori specification and able to handle large amount of noise in dataset. It fails in case of neck type of dataset and it does not work well in case of high dimensionality data [13].

D. Grid Density Based Algorithms

Grid Density based clustering is concerned with the value space that surrounds the data points not with the data points. This algorithm uses the multi resolution grid data structure and use dense grids to form clusters. Grid Density based algorithms require the users to specify a grid size or the density threshold, the problem here arise is that how to choose the grid size or density thresholds. To overcome this problem, a technique of adaptive grids are proposed that automatically determines the size of grids based on the data distribution and does not require the user to specify any parameter like grid size or the density threshold. The grid- based clustering algorithms are STING, Wave Cluster, and CLIQUE

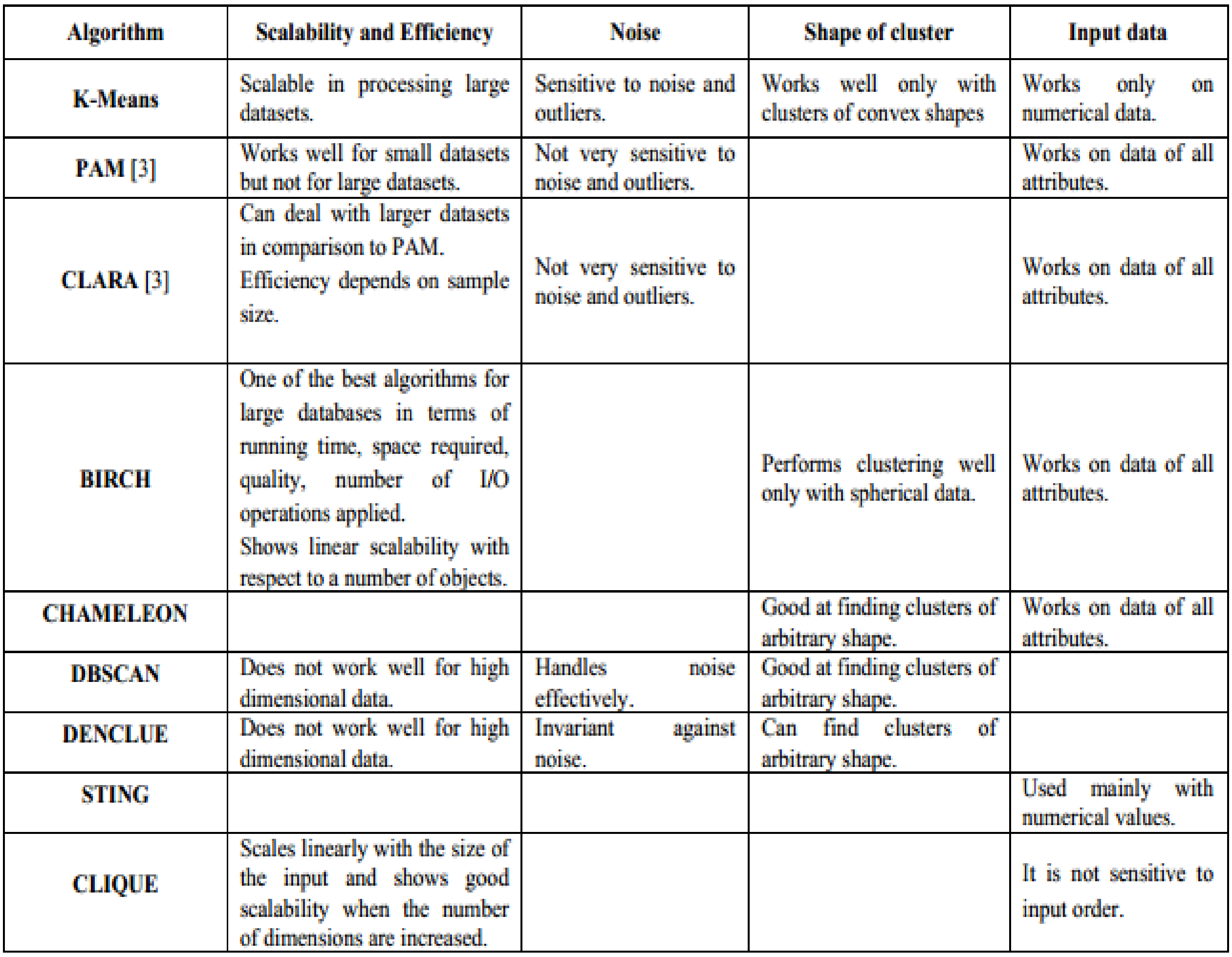
1) STING (Statistical Information Grid approach): This approach breaks the available space of objects into cells of rectangular shapes in a hierarchy. It follows the top down approach and the hierarchy of the cells can contain multiple levels corresponding to multiple resolutions [12]. The statistical information like the mean, maximum and minimum values of the attributes is precomputed and stored as statistical parameters, and is used for query processing and other data analysis tasks. The statistical parameters for higher level cells can be computed from the parameters for lower level cells. The complexity is O (K) where k denotes the total count of cells in last tier.

2) CLIQUE: Clique is a grid-based method that finds density-based clusters in subspaces. CLIQUE performs clustering in two steps. In the first step, CLIQUE partitions each dimension into non-overlapping rectangular units, thereby partitioning the entire space of data objects into cells. At the same time it identifies the dense cells in all the subspaces. The unit is dense when the fraction of total data points exceeds the input model parameter. In the second step, CLIQUE uses these dense cells to form clusters, which can be arbitrary.

Pros and cons of Grid Based Algorithm [7]: The advantage of this algorithm is its quick processing time and this is typically independent of the number of data objects. The Disadvantage is it depends on only the number of cells in each dimension in the quantized space.

IV. CONCLUSION

The objective of the data mining technique is to mine information from a large data set and make it into a reasonable form for the supplementary purpose. Clustering is a significant task in data analysis and data mining applications. It is the task of arranging a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Clustering algorithms can be classified into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. Under partitioning method, a brief description of k-means and k-medoids algorithms have been studied. In hierarchical clustering, the BIRCH and CHAMELEON algorithms have been described. The DBSCAN and DENCLUE algorithms under the density based methods have been studied. Finally, under grid-based clustering method, the STING and CLIQUE algorithms have been described. The challenge with clustering analysis is mainly that different clustering techniques give substantially different results on the same data. Moreover, there is no algorithm present which gives all the desired outputs. Because of this, there is extensive research being carried out in ‘ensembles’ of clustering algorithms, i.e. multiple clustering techniques done on a single dataset.



Algorithm	Scalability and Efficiency	Noise	Shape of cluster	Input data
K-Means	Scalable in processing large datasets.	Sensitive to noise and outliers.	Works well only with clusters of convex shapes	Works only on numerical data.
PAM [3]	Works well for small datasets but not for large datasets.	Not very sensitive to noise and outliers.		Works on data of all attributes.
CLARA [3]	Can deal with larger datasets in comparison to PAM. Efficiency depends on sample size.	Not very sensitive to noise and outliers.		Works on data of all attributes.
BIRCH	One of the best algorithms for large databases in terms of running time, space required, quality, number of I/O operations applied. Shows linear scalability with respect to a number of objects.		Performs clustering well only with spherical data.	Works on data of all attributes.
CHAMELEON			Good at finding clusters of arbitrary shape.	Works on data of all attributes.
DBSCAN	Does not work well for high dimensional data.	Handles noise effectively.	Good at finding clusters of arbitrary shape.	
DENCLUE	Does not work well for high dimensional data.	Invariant against noise.	Can find clusters of arbitrary shape.	
STING				Used mainly with numerical values.
CLIQUE	Scales linearly with the size of the input and shows good scalability when the number of dimensions are increased.			It is not sensitive to input order.

Table 1. Comparison of the features of various clustering algorithms [9]

REFERENCES

[1]. K. Kameshwaran and K. Malarvizhi, “*Survey on Clustering Techniques in Data Mining*”, International Journal of Computer Science and Information Technologies (0975-9646), Vol. 5(2), 2014

[2]. Pradeep Rai and Shubha Singh (2010) *A Survey of Clustering Techniques*, International Journal of Computer Applications (0975 – 8887) Vol 7– No.12, pp. 1-5

[3]. V.Kavitha, M.Punithavalli (2010) *Clustering Time Series Data Stream – A Literature Survey*, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 1, pp. 289-294.

[4]. S. Anitha Elavarasi and Dr. J. Akilandeswari (2011) *A Survey On Partition Clustering Algorithms,* International Journal of Enterprise Computing and Business Systems.

[5]. S.Vijayalaksmi and M Punithavalli (2012) *A Fast Approach to Clustering Datasets using DBSCAN and Applications* (0975 – 8887) Vol 60– No.14, pp. 1-7.

[6]. Cheng-Far Tsai and Tang-Wei Huang (2012) *QIDBSCAN: A Quick Density-Based Clustering Technique* idea International Symposium on Computer, Consumer and Control, pp. 638-641.

[7]. Preeti Baser and Dr. Jatinderkumar R. Saini, *A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets,* International Journal of Computer Science & Communication Networks,Vol 3(4),271-275

[8]. Sunita Jahirabadkar and Parag Kulkarni (2013) *Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms,* International Journal of Computer Applications(0975 – 8887) Vol 63– No.20, pp. 29-35.

[9]. Mihika Shah and Sindhu Nair, *A Survey of Data Mining Clustering Algorithms***,** International Journal of Computer Applications (0975 – 8887) Volume 128 – No.1, October 2015

[10]. Pavel Berkhin, *Survey of Clustering Data Mining Techniques*, Accrue Software, Inc.

[11]. NavneetKaur, *Survey Paper on Clustering Techniques*, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[12]. Nisha and Puneet Jai Kaur, “*A Survey of Clustering Techniques and Algorithms*”, IEEE (978-9-3805-4415-1), 2015

[13]. Han, J. and Kamber, M. *Data Mining- Concepts and Techniques*, 3rd Edition, 2012, Morgan Kauffman Publishers.

[14]. Megha Mandloi, *A Survey on Clustering Algorithms and K-Means*, July-2014

[15]. Amandeep Kaur Mann and Navneet Kaur, “Review Paper on Clustering Techniques”, Global Journal of Computer Science and Technology, Software and Data Engineering (0975-4350), Volume 13 Issue 5 Version 1.0 Year 2013

[16]. Pavel Berkhin, “Survey of Clustering Data Mining Techniques”, Accrue Software, Inc.