A Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining

Amit Kumar Kar[1], Shailesh Kumar Patel[2], Rajkishor Yadav[3]

Department of Computer Science & Engineering, ITM GIDA, Gorakhpur

***Abstract*— The goal of the data mining process is to extract information from a large data set and transform it into an different usable form for further use. Clustering is very important in data analysis and in different data mining applications. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Clustering deals with finding a structure in a collection of unlabeled data. There are different clustering algorithms are used to organize data, categorize data, for data compression and model construction etc. This paper analyzes the four major clustering algorithms namely: Partitioning methods, Hierarchical methods, Grid-based methods and Density-based methods and comparing the performance of these algorithms on the basis of correctly class wise cluster building ability of algorithm.**

***Keywords*—Data mining, Clustering, Hierarchical method, Partitioning method, Grid-based method, Density-based method.**

I. INTRODUCTION

Data mining is the use of automated data analysis methods to uncover the previously undetected relationship among data items [1]. Data mining generally involves the analysis of different data stored in a data warehouse. The most important data mining techniques are classification, clustering and regression. In this research paper we are working on different types of clustering algorithm such as hierarchical methods, partitioning methods, grid-based methods and density-based algorithms.

The important task of clustering is data mining, and a common technique for statistical data analysis used in different fields, with machine learning ability, information retrieval from many sources, different image analysis, pattern recognition, medical research and bioinformatics.

Cluster analysis organizes and summarizes data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups. In cluster analysis the data objects belonging to a cluster are similar, are called homogeneous, and the objects belonging to different clusters, are called heterogeneous. This definition indicates that clustering cannot be a one-step process. Therefore the clustering processes are divided into the following stages [3].

A. *Data Collection:*

It consists of the extraction of relevant and similar data objects from the different data sources. Data objects are differentiated by their own individual values for a set of attributes.

B. *Initial Screening:*

It refers to the messaging of data after its extraction from the sources. This stage is closely connected to a process widely used in Data Warehousing, called Data Cleaning.

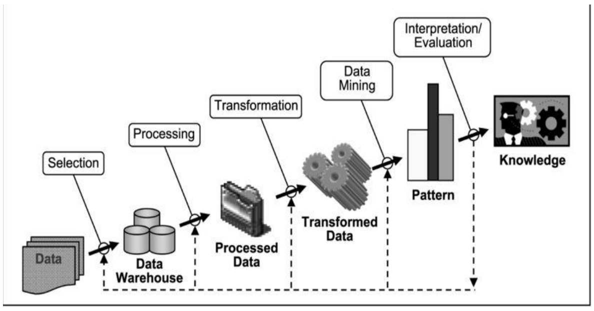


Figure 1: Steps of Data Mining Process

C. *Representation*

It includes the proper preparation of the data in order to become suitable for the clustering algorithm.

D. *Clustering Tendency*

Checks whether the data in hands has a natural tendency to cluster or not. This step is often ignored, especially in the presence of large data set of items.

E. *Clustering Strategy*

It involves the careful considering of clustering algorithm and different values and initial parameters.

F. *Validation*

This is one of the last and most under-studied stages. Validation is based on manual inspection and visual techniques. As the amount of data and their dimensionality will grow, then there is no means to compare and contrast the results with calculated ideas.

G. *Interpretation:*

This stage includes the collection of different clustering results with other studies in order to conclude and further analysis.

II. Data Types And Their Measures

A comprehensive categorization of the different types of variables and conditions met in most data sets provides a helpful means for identifying the differences among data elements.

There are different classifications based on two parameters: the Domain Size and Measurement Scale.

A. Classification Based on the Domain Size

This classification differentiates the data objects based on the size of their domain, i.e, the number of distinct and different values the data objects may assume. In the following discussion we assume a database *D*, of *n* objects or tuples. If $x^{\wedge}$, $y^{\wedge}$ and $z^{\wedge}$ are three data objects belonging to D, each one of them has the form: $x^{\wedge} = (x_1, x_2, \ldots., x_k)$, $y^{\wedge} = (y_1, y_2, \ldots..y_k)$ and $z^{\wedge} = (z_1, z_2, \ldots..,z_k)$ where   is the *dimensionality*, while each x_i, y_i and z_i, $1 \leq i \leq k$, is a *feature*, or an *attribute* of the corresponding object. Hence, the term "data types" will refer to "attribute data types". We have the following classes:

1) An attribute is said to be *continuous*, if its domain value is *uncountably infinite*, *i.e.*, its elements cannot be taken into a one-to-one correspondence with the set of positive integers. It means that between any two values of the attribute, there exist an infinite number of values.
2) An attribute is said to be *discrete*, if its domain is a *finite set*, *i.e.*, a set of elements can be taken into a one-to-one correspondence with a finite subset of the positive integers.

The class of *binary* attributes consists of attributes whose domain includes exactly two discrete values.

B. Classification Based on the Different Scale

This classification shows attributes according to their measurement scales. Suppose we have an attribute R and two tuples $x^{\wedge}$ and $y^{\wedge}$ with values x_i and y_i for this attribute, respectively. Then we have the following classes:

1) A nominal scale means, that we can only say if $x_i = y_i$ or $x_i \neq y_{i.}$ Nominal scaled attribute values cannot be totally ordered. They are just a generalization of binary attributes, with a domain of more than two discrete values.
2) An ordinal scale means nominal scaled attributes with the additional feature that their values can be totally ordered but differences among the scale points cannot be quantified.
3) An interval scale calculates the values in a linear scale. With interval scaling we cannot say only if one value comes before or after another, but also how far before or after.
4) A ratio scale is defined an interval scales with a meaningful zero point.

III. Categorization Of Clustering Methods

Many methods have appeared in order to discover cohesive groups in large datasets. In the following section we present some basic methods or techniques for clustering.

A. Hierarchical

Hierarchical algorithms create a hierarchical decomposition of the objects[4]. Hierarchical clustering builds a cluster Hierarchy that means a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down).

1) *Agglomerative* algorithms

It starts with each object being a separated cluster itself, and successively merging groups according to a distance measure. The clustering may stop when all objects are in a single group or at any other point the user wants. These methods generally follow a greedy-like bottom-up merging.

2) *Divisive algorithms*

It uses the opposite strategy than agglomerative algorithm. It starts with one group of all objects and successively splited the groups into smaller ones, until each object falls in one cluster, or as needed. Divisive approaches divide the data objects in disjoint groups at every step, and follow the same pattern until all objects fall into a separate cluster. This is equivalent to the approach followed by divide-and-conquer algorithms.

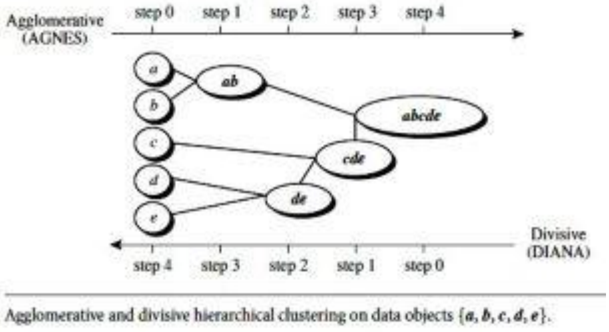


Figure 2: Hierarchical Clustering Algorithms

B. Partitioning

Data partitioning algorithms divides the data into a number of subsets [2]. Because checking all possible subset systems is computationally infeasible, certain greedy heuristics methods are used in the form of iterative optimization technique. Specifically, it means different relocation methods that iteratively reassign points between the k numbers of clusters. Unlike the traditional hierarchical methods, in which clusters are not revisited after being constructed, relocation algorithms gradually improve clusters with appropriate data, this result in high quality clusters.

1) Probabilistic Clustering

In the probabilistic clustering, data is considered to be a sample, independently drawn from a heterogeneous model of different probability distributions. Probabilistic clustering has some important features:

- It can be modified to handle records of complex data structure
- It can be paused with consecutive batches of data, since clusters have different representation which differ from different set of points.
- At any stage of iterative process the intermediate mixture model can be used to assign cases.
- This gives easily interpretable cluster system.

The important property of probabilistic clustering is the mixture model can be generalized to clustering heterogeneous data.

2) K-Medoids Methods

In k-medoids methods, a cluster is shown by one of its points. When medoids are choosen, clusters are defined as subsets of points nearest to respective medoids, and the objective function is termed as the averaged distance measure between a point and its medoid. The different versions of k-medoid methods are PAM (Partitioning Around Medoids) and CLARA (Clustering LARge Applications).

PAM is iterative optimization that combines relocation of points between perspective clusters with re-nominating the points as potential medoids. CLARA uses different samples, each of with 40+2k points, which are each subjected to PAM. The whole dataset is related to resulting medoids, the objective function is calculated, and the best system of medoids is unchanged. Further progress is related to introduced the algorithm CLARANS (Clustering Large Applications based upon RANdomized Search) in the context of clustering in spatial databases. CLARANS uses random search to generate neighbors by starting with an arbitrary node and randomly checking maxneighbor neighbors. If a neighbor indicates a better partition, the process continues with this latest node. Otherwise a local minimum is found, and the algorithm restarts until numlocal local minima are found.

3) K-Means Methods

The k-means method is one of the most popular clustering tools, which used in scientific applications. In this method statistical mean is calculated to compute the k-means. The name comes from representing each of k clusters C_j by the mean (or weighted average) c_j of its point, the so called centroid.

C. Grid-Based

In this method there is formation of number of vertical and horizontal partitions, which called the grid.

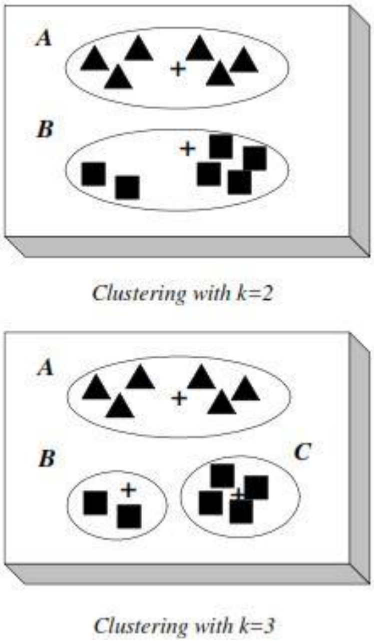


Figure 3: Partitioning Clustering Algorithms

The important focus of this method is spatial data, i.e., data that model the geo- metric structure of objects in space, their relationships, properties and operations. The main objective of these algorithms is to quantize and quantify the data set into the different cells and then the work with objects containing to these cells. They do not relocate points but rather build several hierarchical levels of groups of objects. In this scenario, they are closer to hierarchical algorithms but the merging of grids, and consequently clusters, does not depend on a distance measure but it is decided by a predefined parameter.

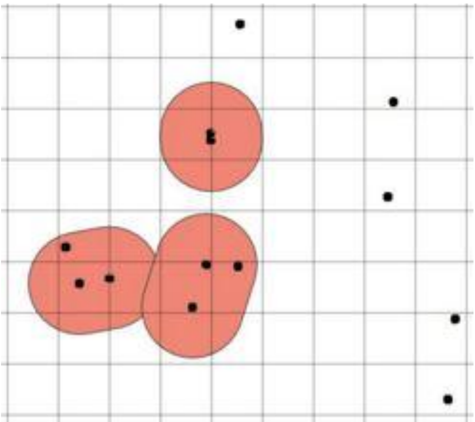
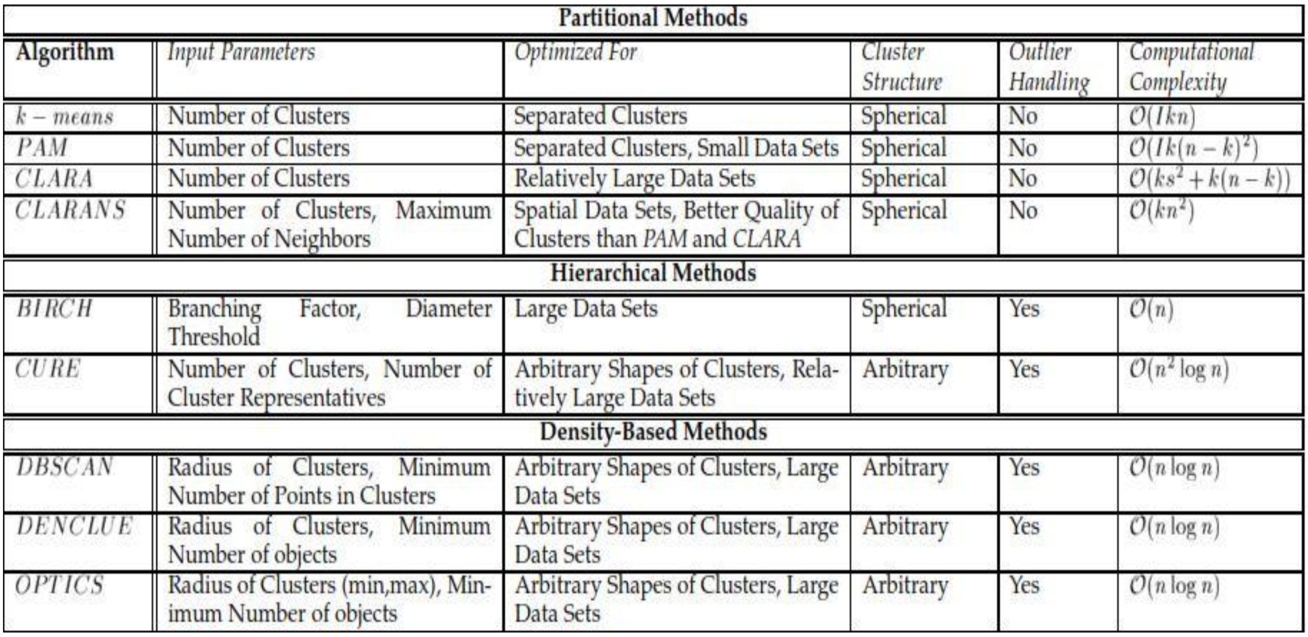


Figure 4: Grid-Based Clustering Algorithms

D. Density-Based

In this method, there are different densities exist on the basis of presence of number of data objects in a specific area. These algorithms group objects according to specific density objective functions [2]. Density is usually defined as the number of objects in a particular neighborhood of a data objects. In this method a given cluster continues raising as long as the number

TABLE 1
COMPARISON AMONG DIFFERENT CLUSTERING ALGORITHM



Algorithm	*Input Parameters*	*Optimized For*	*Cluster Structure*	*Outlier Handling*	*Computational Complexity*
Partitional Methods					
$k-means$	Number of Clusters	Separated Clusters	Spherical	No	$\mathcal{O}(Ikn)$
PAM	Number of Clusters	Separated Clusters, Small Data Sets	Spherical	No	$\mathcal{O}(Ik(n-k)^2)$
$CLARA$	Number of Clusters	Relatively Large Data Sets	Spherical	No	$\mathcal{O}(ks^2+k(n-k))$
$CLARANS$	Number of Clusters, Maximum Number of Neighbors	Spatial Data Sets, Better Quality of Clusters than *PAM* and *CLARA*	Spherical	No	$\mathcal{O}(kn^2)$
Hierarchical Methods					
$BIRCH$	Branching Factor, Diameter Threshold	Large Data Sets	Spherical	Yes	$\mathcal{O}(n)$
$CURE$	Number of Clusters, Number of Cluster Representatives	Arbitrary Shapes of Clusters, Relatively Large Data Sets	Arbitrary	Yes	$\mathcal{O}(n^2 \log n)$
Density-Based Methods					
$DBSCAN$	Radius of Clusters, Minimum Number of Points in Clusters	Arbitrary Shapes of Clusters, Large Data Sets	Arbitrary	Yes	$\mathcal{O}(n \log n)$
$DENCLUE$	Radius of Clusters, Minimum Number of objects	Arbitrary Shapes of Clusters, Large Data Sets	Arbitrary	Yes	$\mathcal{O}(n \log n)$
$OPTICS$	Radius of Clusters (min,max), Minimum Number of objects	Arbitrary Shapes of Clusters, Large Data Sets	Arbitrary	Yes	$\mathcal{O}(n \log n)$

of objects in the neighborhood exceeds some parameters. It is considered to be different from the idea in partitioned algorithms that use iterative relocation of points given a certain number of clusters.

IV. COMPARISON AMONG DIFFERENT CLUSTERING ALGORITHMS

The table 1 has indicated different parameters like structure, complexity of different clustering algorithm in data mining.

V. CONCLUSION

In this research paper, we have discussed different clustering techniques, each of their own pros and cons according to the different situation.

1) In case of hierarchical method, once the process has been completed it cannot be undone. In this method there are different hierarchies between the different data clusters.

2) In case of partitioning method different statistical measures are used like mean, median and mode. It means in this partitioning it is restricted to numeric only.

3) In case of grid-based method, the different grids are constructed of different size.

4) Density-based method is applicable only arbitrary shape but hierarchical and partitioning methods are restricted to spherical shaped cluster.

REFERENCES

[1] Sharmila, R.C.Mishra , "Performance Evaluation of Clustering Algorithms" in International Journal of Engineering Trends and Technology (IJETT)- Volume4 Issue7-July 2013.

[2] Amandeep Kaur Mann and Navneet Kaur, “Survey Paper on Clustering Techniques”, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[3] Periklis Andritsos,”Data Clustering Techniques-Qualifying Oral Examination Paper”,University of Toronto, Department of Computer Science, March 11, 2002.

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