This is to certify that Prof. / Dr. / Mr. / Ms. Manoj Kr. Gupta of RDIAS, Delhi has attended/contributed/presented a paper entitled "A Comparative Study of Clustering Algorithms" during INDIACom-2019; 13th INDIACom; 2019 6th International Conference on "Computing for Sustainable Global Development" organised by BVICAM, New Delhi.

(Prof. A. K. Saini)
Conference Chair

(Prof. M. N. Hoda)
General Chair
A Comparative Study of Clustering Algorithms

Manoj Kr. Gupta
RDIAS, Delhi, INDIA.
USIC&T, GGSIPU, New Delhi, INDIA
Email ID: manojkgupta5@gmail.com

Pravin Chandra
USIC&T, GGSIPU, Delhi, INDIA
Email ID: pravin.chandra.pravin@gmail.com

Abstract—In present era, data analysis plays vital role in various domains. Data clustering is a data analysis technique used for grouping of data objects based on unsupervised learning. Many clustering algorithms have been proposed in the literature. Each algorithm possesses some strengths and weaknesses. Therefore, a set of clustering algorithms are appropriate for one set of application area while another set of clustering algorithm are suitable for another set of application areas. In this paper, popular traditional algorithms are discussed. A comprehensive comparative study of different clustering algorithms is presented in this paper. These clustering algorithms are compared in detail based on various parameters used in these methods.

Keywords—Clustering Algorithms, Cluster Analysis, Comparative Study, Survey, Data Mining.

I. INTRODUCTION

Clustering is a vital exploratory multivariate data analysis method. In clustering, data objects are partitioned into clusters based on distance / dissimilarity among data objects. Data objects which are like or near to each other are placed within the same cluster while unlike or far off data objects are placed in another cluster. Like classification, clustering is also classifying the data objects but unlike the classification, the class labels are unknown because clustering is based on unsupervised learning [16]. The clusters are defined based on the study of the behaviour or characteristics of the data objects by the domain experts.

The clustering algorithms must have following properties:

a) Data objects within the cluster must be like or near to each other as much as possible.

b) Data objects belong to different clusters must be dissimilar or far off to each other as much as possible.

c) The distance / similarity measure must have some practical ability and be clear.

Clustering is also extensively used in many application domains i.e. statistics, image segmentation, image pattern recognition, object recognition, information retrieval, bioinformatics, etc. [21].

Distance and similarity is the foundation for formulation of clusters by the clustering algorithms. Distance measures are preferred for quantitative data whereas similarity measures are preferred for qualitative data [50]. Different distance metrics and similarity measures have been proposed by the many researchers which can be useful for clustering. Though, Euclidean, Manhattan and Mahalanobis metrics are widely used in different clustering algorithms.

The clustering process is presented in section 2. Comparison among various clustering algorithms is systematically analyzed and presented in section 3. Finally, conclusion is drawn in section 4.

II. CLUSTERING PROCESS

A typical data clustering process includes several steps which are presented in Fig. 1. In the following subsections, these steps are described [15, 21]:

A. Feature selection

It is a data preprocessing step based on dimensionality reduction. In this step, the appropriate features or characteristics of data are selected. Irrelevant features will leads to the unnecessary complexity in the clustering process.

B. Clustering algorithm

For clustering, either an appropriate existing cluster algorithm is chosen or a clustering algorithm is designed as per the required clustering goal. The chosen algorithms will be executed to find out the clusters within the data set. The appropriate clustering algorithm will generate better results as compared to the other algorithms.

C. Validation of the results

The identified clusters need to be validated as to know whether the clusters are identified appropriately or not. There are three categories of cluster validation techniques: (a) internal (b) external and (c) relative.

D. Interpretation of the results

After validation of the clusters, the clusters needs to analyzed and interpreted to describe the each cluster. The interpretation is to draw the right conclusion.

Fig. 1. Steps of a Typical Clustering Process
III. COMPARATIVE STUDY OF CLUSTERING ALGORITHMS

A lot of data clustering algorithms has been proposed by the many researchers so far. Some of these algorithms are based on one method / approach while other algorithms are based on another method / approach. Based on the similarity of method / approach followed by the algorithms, the clustering algorithms can be generally classified into the following categories [21, 16, 11, 45]:

A. Hierarchical Method

A hierarchical decomposition of the sets of data objects is identified and created by hierarchical method. Based on the order of hierarchical decomposition, these methods can be classified as either agglomerative or divisive. SLINK [40], CLINK [18], BIRCH [53], CURE [13], ROCK [14] algorithms follow agglomerative approach whereas DIANA [27] and DISMEXA [43] algorithms follow divisive approach. Hierarchical methods can be based on either distance or density or continuity. In these methods, if merge or split is done then it can never be undone. In other words, these cannot correct erroneous decisions. The comparison of various popular clustering algorithms based on hierarchical method is presented in Table I. Some more hierarchical algorithms can be found in [20, 41, 46, 50, 58, 59, 70, 71, 72, 73].

B. Partition-based Method

A partition-based method constructs predetermined k partitions (or clusters) of a data set of n objects, where k ≤ n. k-Means [28], k-Medoids [30], PAM [26], CLARA [25] and CLARANS [29] algorithms are based on partitioning methods. These methods are effective for data sets up to medium size. But, these can be extended for the large data sets as well. The comparison of various popular clustering algorithms based on partitioned-based method is presented in Table II. Some more partition-based algorithms can be found in [20, 26, 33, 38, 41, 47, 50, 55, 56, 57, 70, 71, 72, 73].

C. Density-based Method

In these methods, objects are clustered based on the concept of density instead of distance. Hence, arbitrary-shaped clusters can also be formed by these methods. The idea behind these methods is to keep on increasing a given cluster provided that the density in the neighborhood exceeds some threshold. These may also filter out outliers. DBSCAN [9], OPTICS [2], DENCLUE [17] and RDBC [44] algorithms are the examples of density-based methods. The comparison of various popular clustering algorithms based on density-based method is presented in Table III. Some more density-based algorithms can be found in [20, 22, 34, 41, 50, 63, 64, 65, 66, 70, 71, 72, 73].

D. Grid-based Method

These methods quantize the object space into a grid structure. These methods are fast because of independence from number of data objects, yet reliant on grid size in each dimension in the quantized tree. STING [48], CLIQUE [1], OptiGrid [18], GRIDCLUS [36], GDILC [54], WaveCluster [39] are grid-based methods. The comparison of various popular clustering algorithms based on grid-based method is presented in Table IV. Some more grid-based algorithms can be found in [20, 41, 42, 50, 67, 68, 69, 70, 71, 72, 73].

E. Fuzzy-based Method

Unlike other crisp methods, a fuzzy-based method assigns the each data object to all the clusters with definite degree of membership. Mostly partitioned-based methods are extended for fuzzy clusters. Fuzzy k-means [3], FCM [4], FCS [7], MM [51], MEC [35] are fuzzy-based clustering algorithms. The comparison of various popular clustering algorithms based on fuzzy-based method is presented in Table V. Some more fuzzy-based algorithms can be found in [20, 41, 50, 60, 61, 62, 70, 71, 72, 73].

F. Modern Clustering Methods

Apart from the above-mentioned categories of the clustering algorithms, the modern clustering algorithms are also classified into some more categories as [16, 50, 70, 72, 73] (a) Kernel-based algorithms, (b) Ensemble-based algorithms, (c) nature-inspired algorithms, (d) graph-based algorithms, (e) model-based algorithms, (f) Quantum Theory-based algorithms, etc.

IV. SUMMARY AND CONCLUSION

The paper started with the introduction of clustering followed by typical clustering process. A comprehensive comparative study of traditional clustering algorithms is presented in Table I through Table V. The main objective of this paper is to present the main idea of the common clustering algorithms and compare them based on several parameters. Only some common popular algorithms are discussed in this paper due to difficult to study and present all proposed algorithms because of the availability of a large number of clustering algorithms in a large number of sources. However, a systematic and clear view of the important algorithms is presented in this paper.

REFERENCES

A Comparative Study of Clustering Algorithms

of the second ACM SIGKDD international conference on knowledge discovery and data mining, pp 226–231
of the 1996 ACM SIGMOD international conference on management of
[54] Zhao, Y. and Song, J.(2001), GDILC: A grid-based density-isoline
clustering algorithm. In Proceedings of the international conferences on
Mohabier B, Moore J (1999), Partitioning-based clustering for web
Recognit Lett 31:651–666
clustering algorithms in data mining: an experimental approach. Inf
Technol J 10:478–484
analysis and image recognition. Wiley, Hoboken
clustering. In: Proceedings of the 8th european conference on principles
and practice of knowledge discovery in databases, pp 231–244

<table>
<thead>
<tr>
<th>Algorithm Basis</th>
<th>SLINK</th>
<th>CLINK</th>
<th>BIRCH</th>
<th>CURE</th>
<th>ROCK</th>
<th>Chameleon</th>
<th>DIANA</th>
<th>DISMEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>[40]</td>
<td>[18]</td>
<td>[53]</td>
<td>[13]</td>
<td>[14]</td>
<td>[23]</td>
<td>[27]</td>
<td>[43]</td>
</tr>
<tr>
<td>Algorithm Type</td>
<td>Agglomerative</td>
<td>Agglomerative</td>
<td>Agglomerative</td>
<td>Agglomerative</td>
<td>Agglomerative</td>
<td>Agglomerative</td>
<td>Divisive</td>
<td>Divisive</td>
</tr>
<tr>
<td>Method</td>
<td>Graph</td>
<td>Graph</td>
<td>Geometric</td>
<td>Geometric</td>
<td>Geometric</td>
<td>Graph</td>
<td>Gini Index</td>
<td>k-means</td>
</tr>
<tr>
<td>Representation</td>
<td>Pointer Representation of Dendrogram</td>
<td>Cluster feature vector</td>
<td>Fixed number of points</td>
<td>links</td>
<td>k-nearest neighbor graph</td>
<td>Series of Successive splits in Dendrogram or Banners</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Carried out using</td>
<td>Arbitrary dissimilarity coefficient</td>
<td>Dissimilarity measure</td>
<td>CF tree</td>
<td>Combinator of random sampling and partitioning</td>
<td>Computations on links</td>
<td>Construction of Sparse Graph</td>
<td>Largest dissimilarity</td>
<td>k-means algorithm to subdivide a cluster</td>
</tr>
<tr>
<td>Type of Data</td>
<td>Numerical / Categorical</td>
<td>Numerical / Categorical</td>
<td>Numerical</td>
<td>Numerical</td>
<td>Boolean and Categorical</td>
<td>Numerical / Categorical</td>
<td>Numerical / Categorical</td>
<td>Categorical</td>
</tr>
<tr>
<td>Shape of Clusters</td>
<td>Arbitrary</td>
<td>Convex</td>
<td>Convex with uniform size</td>
<td>Arbitrary with wide variance in size</td>
<td>Arbitrary with wide variance in size</td>
<td>Arbitrary shape</td>
<td>Convex with uniform size</td>
<td>Convex</td>
</tr>
<tr>
<td>Sensitive to Outliers / Noise</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Scalability for Large Data Sets</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Scalability for High Dimensionality</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Time Complexity</td>
<td>O(n²)</td>
<td>O(n²)</td>
<td>O(n)</td>
<td>O(n log n)</td>
<td>O(n log n)</td>
<td>O(n²)</td>
<td>O(n log n)</td>
<td>O(n²)</td>
</tr>
</tbody>
</table>

TABLE I. COMPARISON AMONG CLUSTERING ALGORITHMS BASED ON HIERARCHICAL METHOD

Proceedings of the 13th INDIACom; INDIACom-2019; IEEE Conference ID: 46181

Proceedings of the eleventh ACM SIGKDD international conference on
knowledge discovery in data mining, pp 672–677
In: Proceedings of the 13th ACM SIGKDD international conference on
knowledge discovery and data mining, pp 133–142
clustering algorithm with noise. Inf Syst 32:978–986
Recognit 37:503–514
ACMSIGMOD Rec 33:32–37
high-dimensional very large spatial data bases. Pattern Recognit Lett
26:999–1010
Clustering Algorithms in Data Mining Technique, International Journal
Data Mining—A Survey, IETE Journal of Research, 47:1-2, 19-28, DOI:
10.1080/03772063.2001.11416199
Algorithms and Applications, International Journal of Information
Retrieval Research, 3(2), 1-293
Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional
Data. Springer, Berlin, Heidelberg

A Comparative Study of Clustering Algorithms

### TABLE II. COMPARISON AMONG CLUSTERING ALGORITHMS BASED ON PARTITIONED-BASED METHOD

<table>
<thead>
<tr>
<th>Basis</th>
<th>Algorithm</th>
<th>K-means</th>
<th>K-medoids</th>
<th>PAM</th>
<th>CLARA</th>
<th>CLARANS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>[28]</td>
<td>[30]</td>
<td>[26]</td>
<td>[25]</td>
<td>[29]</td>
<td></td>
</tr>
<tr>
<td>Type of Data</td>
<td>Numerical</td>
<td>Numerical / Categorical</td>
<td>Numerical</td>
<td>Numerical</td>
<td>Numerical</td>
<td></td>
</tr>
<tr>
<td>Shape of Clusters</td>
<td>Convex</td>
<td>Convex</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td></td>
</tr>
<tr>
<td>Result</td>
<td>Centre of Clusters</td>
<td>Medoids of Clusters</td>
<td>Medoids of Clusters</td>
<td>Medoids of Clusters</td>
<td>Medoids of Clusters</td>
<td></td>
</tr>
<tr>
<td>Sensitive to Outliers / Noise</td>
<td>More</td>
<td>Less</td>
<td>Less</td>
<td>Less</td>
<td>Less</td>
<td></td>
</tr>
<tr>
<td>Scalability for Large Data Sets</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Scalability for High Dimensionality</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Time Complexity</td>
<td>$O(kn)$</td>
<td>$O(k(n-k)^2)$</td>
<td>$O(k(n-k)^2)$</td>
<td>$O(k^2+k(n-k))$</td>
<td>$O(k^n)$</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE III. COMPARISON AMONG CLUSTERING ALGORITHMS BASED ON DENSITY-BASED METHOD

<table>
<thead>
<tr>
<th>Basis</th>
<th>Algorithm</th>
<th>DBSCAN</th>
<th>OPTICS</th>
<th>Mean-shift</th>
<th>DENCLUE</th>
<th>RDBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>[9]</td>
<td>[2]</td>
<td>[6]</td>
<td>[17]</td>
<td>[44]</td>
<td></td>
</tr>
<tr>
<td>Shape of Clusters</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td></td>
</tr>
<tr>
<td>Sensitive to Outliers / Noise</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Low</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Scalability for Large Data Sets</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>No of parameters</td>
<td>Small</td>
<td>Large</td>
<td>Small</td>
<td>Large</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>Time Complexity</td>
<td>$O(n \log n)$</td>
<td>$O(n \log n)$</td>
<td>$(\text{kernel})$</td>
<td>$O(n^2)$ or $n \log n$</td>
<td>$O(n^2)$</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE IV. COMPARISON CLUSTERING ALGORITHMS BASED ON GRID-BASED METHOD

<table>
<thead>
<tr>
<th>Basis</th>
<th>Algorithm</th>
<th>STING</th>
<th>CLIQUE</th>
<th>OptiGrid</th>
<th>GRIDCLUS</th>
<th>GDILC</th>
<th>WaveCluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>[48]</td>
<td>[1]</td>
<td>[18]</td>
<td>[36]</td>
<td>[54]</td>
<td>[39]</td>
<td></td>
</tr>
<tr>
<td>Shape of Clusters</td>
<td>Arbitrary</td>
<td>Convex</td>
<td>Arbitrary</td>
<td>Convex</td>
<td>Arbitrary</td>
<td>Arbitrar y</td>
<td></td>
</tr>
<tr>
<td>Sensitive to Outliers / Noise</td>
<td>Low</td>
<td>Moderate</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Scalability for Large Data Sets</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Scalability for High Dimensionality</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Time Complexity</td>
<td>$O(n)$</td>
<td>$O(n+k^2)$</td>
<td>Between $O(nd)$ and $O(d n \log n)$</td>
<td>$O(n)$</td>
<td>$O(n)$</td>
<td>$O(n)$</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE V. COMPARISON CLUSTERING ALGORITHMS BASED ON FUZZY METHOD

<table>
<thead>
<tr>
<th>Basis</th>
<th>Algorithm</th>
<th>Fuzzy k-means</th>
<th>Fuzzy k-modes</th>
<th>FCM</th>
<th>FCS</th>
<th>MM</th>
<th>MEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>[3]</td>
<td>[19]</td>
<td>[4]</td>
<td>[7]</td>
<td>[51]</td>
<td>[35]</td>
<td></td>
</tr>
<tr>
<td>Carried out using</td>
<td>Fuzzy k-partition of data set and k-prototypes</td>
<td>Fuzzy k-partition</td>
<td>Nondegenerate fuzzy c-partitions</td>
<td>Fuzzy scatter matrix</td>
<td>Graph</td>
<td>Statistical physics; Maximizing the entropy at a given average variance</td>
<td></td>
</tr>
<tr>
<td>Shape of Clusters</td>
<td>Convex</td>
<td>Convex</td>
<td>Convex</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Convex</td>
<td></td>
</tr>
<tr>
<td>Sensitive to Outliers / Noise</td>
<td>Yes</td>
<td>No</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>Type of Data</td>
<td>Numerical</td>
<td>Numerical / Categorical</td>
<td>Numerical</td>
<td>Numerical</td>
<td>Numerical</td>
<td>Numerical</td>
<td></td>
</tr>
<tr>
<td>Time Complexity</td>
<td>$O(kn)$</td>
<td>$O(k(n-k)^2)$</td>
<td>$O(n)$</td>
<td>$(\text{kernel})$</td>
<td>$O(n^2)$</td>
<td>$O(n)$</td>
<td></td>
</tr>
</tbody>
</table>