Volume 5, Issue 2, February 2020 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

A Recent Comparative Survey on Various Clustering Techniques

Tarrnum Khan
P.G. Research Scholar
C.S.E. Department
JIT Borawan Khargone

Mr. Ranjan Thakur
Asst. Professor
C.S.E. Department
JIT Borawan Khargone

Abstract:-Clustering is one of the most common empirical data analysis techniques. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. Data points in each cluster are as similar as possible according to a similarity measure such as Euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application specific. Clustering analysis can be done on the basis of features try to find subgroups of samples based. Cover here clustering based on features. Clustering is used in market segmentation; customers that are similar to each other whether in terms of behaviours or attributes, image segmentation/compression; to group similar regions together, document clustering based on topics, etc. In this paper we proposed a recent comparative study of various clustering techniques

Keywords: Keywords:- Clustering, Distance, Similarity Sub groups, Features

I. INTRODUCTION

A cluster is a subset of objects which are "similar". A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it.
A good clustering method will produce high quality clusters in which:
•The intra-cluster similarity is high.
•The inter-class similarity is low.
The quality of a clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

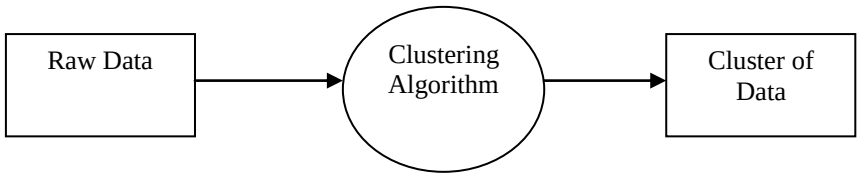

Figure 1 Clustering process

REQUIREMENTS OF CLUSTERING

The following points throw light on why clustering is required

1. *Scalability* −We need highly scalable clustering algorithms to deal with large databases.
2. *Ability to deal with different kinds of attributes* − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
3. *Discovery of clusters with attribute shape* − the clustering detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
4. *High dimensionality* − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
5. *Ability to deal with noisy data* − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
6. *Interpretability* −The clustering results should be interpretable, comprehensible, and usable.

APPLIACTIONS OF CLUSTERS ANALYSIS

There are several application where clustering is used some of them are

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster

TYPES OF CLUSTERING

Clustering methods are divided into following categories

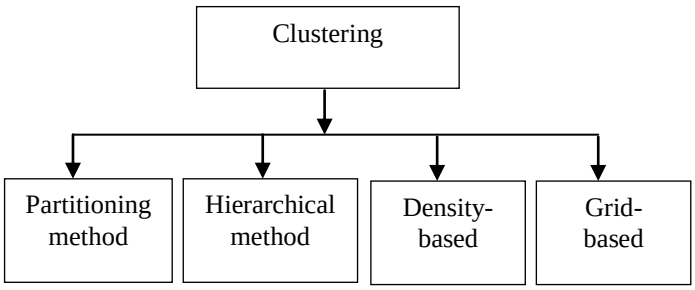

Figure 2 various clustering methods

Partition based clustering are dividing into 3 category

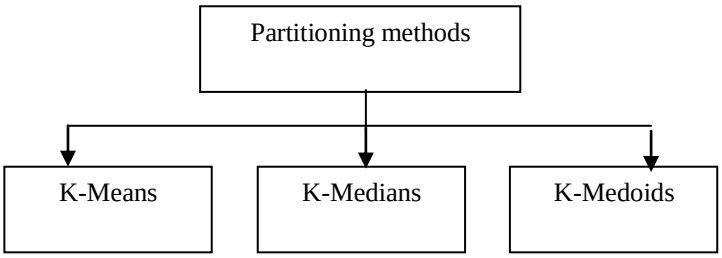

Figure 3 partition based clustering methods

Hierarchical based clustering are dividing into 3 category

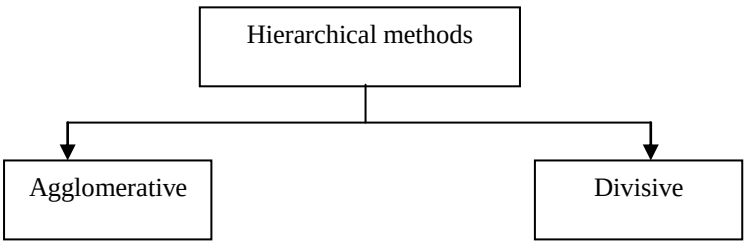

Figure 4 Hierarchical based clustering methods

II. LITERATURE SURVEY

In 2011 Mikko Malinen et al. "K-means Clustering by Gradual Data Transformation". Traditional approach to clustering is to fit a model for the given data. They proposed a completely opposite approach by fitting the data into a given clustering model that is optimal for similar pathological data of equal size and dimensions. They performed inverse transform from this synthetic data back to the original data while refining the optimal clustering structure during the process. The key idea is that we do not need to find optimal global allocation of the prototypes. They only need to perform local fine-tuning of the clustering prototypes during the transformation in order to

preserve the already optimal clustering structure They proposed an alternative approach for clustering by fitting the data to the clustering model and not vice versa [1].

In 2012 P. Indira Priya et al. "K-means Clustering Algorithm Characteristics Differences based on Distance Measurement". They proposed a new Minkowski distance based K-means algorithm called Enhanced K-means Clustering algorithm (EKMCA) is proposed and also demonstrates the effectiveness of the distance measurement, the performance of this kind of distance and the Euclidian and Minkowski distances were compared by clustering KDD'99 Cup dataset. Experiment results show that the new distance measure can provide a more accurate feature model than the classical Euclidean and Manhattan distances. The proposed algorithm achieves the high performance when compared with K – means clustering algorithm which is used Euclidean distance measurement [2].

In 2012 Youguo Li, Haiyan Wu "A Clustering Method Based on K-Means Algorithm". They combine the largest minimum distance algorithm and the traditional K-Means algorithm to propose an improved K-Means clustering algorithm. This improved algorithm can make up the shortcomings for the traditional K-Means algorithm to determine the initial focal point. According to the academic analysis and result of experiment above, the improved K-Means not only keeps the high efficiency of standard K-Means but also raises the speed of convergence effectively by improving the way of selecting initial cluster focal point [3].

In 2012 Shailendra Singh Raghuwanshi "Comparison of K-means and Modified K-mean algorithm for Large Data set". They proposed a Modified approach K-Means clustering which executes K-means algorithm this Algorithm approach is better in the process in large number of clusters and its time of execution is comparisons base on K-Mean approach. If the process experimental result is using the proposed algorithm it time of computation can be reduced with a group in runtime constructed data sets are very promising. Modified Approach of K Mean Algorithm is better than K Mean for Large Data Sets. Clustering Efficient K-means algorithm based on iterative process. From the experimental results, it is analysis in the comparison between K-mean and Modified approach K-mean algorithm shows that when it based on the number of records is less, Modified approach K mean takes minimum time to execute than the K-mean[4].

In 2013 Laurence Morissette et al. "The k-means clustering technique: General considerations and implementation in Mathematical". Data clustering techniques are valuable tools for researchers working with large databases of multivariate data. They presented a simple yet powerful one: the *k*-means clustering technique, through three different algorithms: the Forgy /Lloyd, algorithm, the Mac Queen algorithm and the Hartigan & Wong algorithm. They implementation in Mathematical and various examples of the different options available to illustrate the application of the technique. They showed that *k*-means clustering is a very simple and elegant way to partition datasets [5].

In 2014 Cosmin Marian et al. "An Optimized Version of the K-Means Clustering Algorithm". They introduced an optimized version of the standard K-Means algorithm. The optimization refers to the running time and it comes from the observation that after a certain number of iterations, only a small part of the data elements change their cluster, so there is no need to re-distribute all data elements. The prototype implementation showed up to 70% reduction of the running time. They optimized version of the K-Means algorithm. The optimization refers to the running time. Optimization comes from the considerable reduction of the data space that is re-visited at each loop. The data set has been generated using a uniform distribution generator [6].

In 2014 Dr. Manju Kaushik, Mrs. Bhawana Mathur "Comparative Study of K-Means and Hierarchical Clustering Techniques" They compare with k-Means Clustering and Hierarchical Clustering Techniques. Strength and weakness of both Clustering Techniques and their methodology and process.The performance of K- mean algorithm is better than Hierarchical Clustering Algorithm. Performance of K-Means algorithm increases as the RMSE decreases and the RMSE decreases as the number of cluster increases [7].

In 2015 Ramzi A. Haraty, Mohamad Dimishkieh, "An Enhanced *k*-Means Clustering Algorithm for Pattern Discovery in Healthcare Data". They presented a study over data mining applications in healthcare. Mainly, the study of *k*-means clustering algorithms on large datasets and present an enhancement to *k*-means clustering, which requires *k* or a lesser number of passes to a dataset. The proposed algorithm, which we call *G*-means, utilizes a greedy approach to produce the preliminary centroids and then takes *k* or lesser passes over the dataset to adjust these center points. The experiments also yield better results for *G*-means in terms of the coefficient of variance and the execution time [8].

In 2015 Min Wei et al."Clustering Heterogeneous Data with *k*-Means by Mutual Information-Based Unsupervised Feature Transformation". They integrated the MI-based UFT which can transform non-numerical features into numerical features with the conventional *k*-means to cluster the heterogeneous data. The transformation of UFT is based on MI and can preserve the information contained in the original non-numerical features. The results of simulation studies show that, the integrated UFT-*k*-means outperformed other clustering algorithms and provided reasonable clusters for one modified real-world dataset and five real-world benchmark datasets [9].

In 2015 Zeynel Cebeci et al. "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures" They compared K-means (KM) and the Fuzzy C-means (FCM) algorithms for their computing performance and clustering accuracy on different shaped cluster structures which are regularly and irregularly scattered in two an enhancement to *k*-means clustering, which requires *k* or a lesser number of passes to a dataset. The proposed algorithm, which we call *G*-means, utilizes a greedy approach to produce the preliminary centroids and then takes *k* or lesser passes over the dataset to adjust these center points. Experimental results, which were used in an increasing manner on the same dataset, show that *G*-means outperforms *k*-means in terms of entropy and *F*-scores. [10].

In 2016 Paul Inuwa Dalatu "Time Complexity of K-Means and K-Median Clustering Algorithm in Outliers Detection". Data mining responsibilities show the broad features of the data in the database and also examine the current data in order to determine some arrangements. While, clustering establishes an important part of professed data mining, a procedure of exploring and analyzing large volumes of data in order to determine valuable information. Outliers are points that do not conform to the common performance of the data. Therefore, in errors minimization, the K-Medians clustering algorithm is more effective probably because it uses median as robust for computing the clustering compare to the K-Means which uses mean which is sensitive to outliers [11].

In 2016 Unnati R. Raval et al. "Implementing & Improvisation of K-means Clustering Algorithm". The main purpose of the article is to proposed techniques to enhance the techniques for deriving initial centroids and the assigning of the data points to its nearest clusters. They proposed clustering techniques for enhancing the accuracy and time complexity but it still needs some further improvements and in future it is also viable to include efficient techniques for selecting value for initial clusters (k). They explains the techniques that improves the techniques for determining initial centroids and assigning data points to its nearest clusters with more accuracy with time complexity of O(n) which is faster than the traditional k-means. [12].

In 2016 Pooja Pandey et al."Comparison between Standard K-Mean Clustering and Improved K-Mean Clustering". They discussed both phases have some shortcomings and two methods are purposed based on that. First one is about how to generate the centroids and the second one will reduce the time while calculating distance from centroid. K-means is an algorithm typically used for clustering large data sets. They elaborates k-means algorithm, analyze its shortcomings and also purposes the alternatives to these shortcomings. They explains the simple and efficient way of assigning centroids to clusters using cosine, Sorensen dice and Manhatten distance formula and an improved version of k-means which ensures the entire process in O(nk) time without altering the accuracy of clusters[13].

In 2017 Arpit Bansal et al. "Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining" Clustering is technique which is used to analyze the data in efficient manner and generate required information. Central points are selected using the formulae Euclidian distance and on the basis of Euclidian distance points are assigned to the clusters. They proposed an improvement in the k-mean clustering algorithm will be proposed which can define number of clusters automatically and assign required cluster to un-clustered points. The proposed improvement will leads to improvement in accuracy and reduce clustering time by the member assigned to the cluster to predict cancer.[14].

In 2017 Xuanxia Yaoa et al. "An Improved Clustering Algorithm and Its Application in We Chat Sports Users Analysis". They proposed an appropriate method to update clusters number. Although many researchers have been made for numerical, categorical or mixed datasets, most of them are not very effective or cannot guarantee the unique clustering result. To address these problems, an improved clustering algorithm based on entropy is put forward, which uses the divergence to determine the initial cluster centers and introduce the inter-cluster entropy for mixed data to update clusters number. The experiments on the 3 dataset in UCI and the practical dataset of We Chat sports users show that the improved algorithm can guarantee the unique clustering result and has good performance [15]

In 2018 R. Indhu, R. Porkodi "Comparison of Clustering Algorithm". They proposed comparative analysis of four clustering algorithms namely K-means algorithm, Hierarchical algorithm and Density based algorithm and Expectation maximization algorithm. These algorithms are compared in terms of efficiency and accuracy and observed that K-means produces better results as compared to other algorithms. They proposed comparative study has been performed on the analysis of four clustering algorithms: k-means, hierarchical, density based and expectation maximization clustering algorithms [16].

III. COMPARATIVE ANALYSIS

S. No	Clustering techniques	Advantages	Disadvantages
1	Partitioning Clustering	Relatively scalable and simple	Poor cluster descriptors
2	Hierarchical Clustering	No need to define number of clusters in advance.	Inability to make corrections once the splitting/merging decision is made
3	Density Based Clustering	It handles noise and outliers efficiently	Unsuitable for high-dimensional datasets
4	Grid-based Clustering	It is fast as there is no distance computation	All the clusters boundaries are either horizontal or vertical and no diagonal boundary is detected.
5	Model-based Clustering	Clusters can be characterized by a small number of parameters	Computationally expensive

REFERENCE

1. Mikko Malinen and Pasi Fr¨anti "K-means: Clustering by Gradual Data Transformation" 2011 Sixth International Conference on Image and Graphics Speech and Image Processing Unit, School of Computing, University of Eastern Finland.

2. P.Indira Priya " K-means Clustering Algorithm Characteristics Differences based on Distance Measurement" International Journal of Computer Applications (0975 – 8887) Volume 59– No.14, December 2012
3. Youguo Li, Haiyan Wu "A Clustering Method Based on K-Means Algorithm" 2012 International Conference on Solid State Devices and Materials Science © 2012 Published by Elsevier B.V. Selection and/or peer-review under responsibility of Garry Lee.
4. Shailendra Singh Raghuwanshi "Comparison of K-means and Modified K-mean algorithms for Large Data-set" Volume 1, No.3, November – December 2012 International Journal of Computing, Communications and Networking Available Online at http://warse.org/pdfs/ijccn02132012.pdf.
5. Laurence Morissette and Sylvain Chartier "The k-means clustering technique: General considerations and implementation in Mathematica" Tutorials in Quantitative Methods for Psychology 2013, Vol. 9(1), p. 15-24.
6. Cosmin Marian Poteras, "An Optimized Version of the K-Means Clustering Algorithm" Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 695–699 DOI: 10.15439/2014F258 ACSIS, Vol. 2
7. Dr. Manju Kaushik, Mrs. Bhawana Mathur "Comparative Study of K-Means and Hierarchical Clustering Techniques" international journal of software and hardware research in engineering"issue 2 volume 6 june 2014
8. Ramzi A. Haraty,1 Mohamad Dimishkieh "An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data" Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Article ID 615740 International Journal of Distributed Sensor Networks.
9. Min Wei , Tommy W. S. Chow and Rosa H. M. Chan "Clustering Heterogeneous Data with k-Means by Mutual Information-Based Unsupervised Feature Transformation" Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong; E-Mails: eetchow@cityu.edu.hk (T.W.S.C.); rosachan@cityu.edu.hk (R.H.M.C.) Entropy 2015, 17, 1535-1548; doi:10.3390/e17031535.
10. Zeynel Cebeci1, Figen Yildiz "Fuzzy C-Means Algorithms on Different Agricultural Cluster Structures" Journal of 862X) 2015 Vol. 6, No. 3:13-23 Hungarian Association of Agricultural Informatics European Federation for Information Technology in Agriculture, Food and the Environment
11. Paul Inuwa Dalatu "Time Complexity of K-Means and K-Medians Clustering Algorithms in Outliers Detection" Global Journal of Pure and Applied Mathematics. ISSN 0973-1768 Volume 12, Number 5 (2016), pp. 4405–4418 © Research India Publications http://www.ripublication.com/gjpam.htm.
12. Unnati R. Raval, Chaita Jani "Implementing & Improvisation of K-means Clustering Algorithm" Unnati R. Raval et al, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 191-203 .
13. Pooja Pandey "Comparison between Standard K-Mean Clustering and Improved K-Mean Clustering" International Journal of Computer Applications (0975 – 8887) Volume 146 – No.13, July 2016.
14. Arpit Bansal " Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining" International Journal of Computer Applications (0975 – 8887) Volume 157 – No 6, January 2017.
15. Xuanxia Yaoa, Shuying "An Improved Clustering Algorithm and Its Application in We Chat Sports Users Analysis" 2017 International Conference on Identification, Information and Knowledge in the Internet Procedia Computer Science 129 (2018) 166–174
16. R. Indhu, R. Porkodi "Comparison of Clustering Algorithm" International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307.

.