ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

A SURVEY ON CLUSTER ANALYSIS

Mohammed Fasi Ahmed Parvez[*1] **& G. Pradeepini**[2]
[*1]Research Scholar- KL University
[2]Professor- KL University

ABSTRACT

Clustering has become an increasingly popular method of multivariate analysis over the past two decades, and with it has come a vast amount of published material. Since there is no journal devoted exclusively to cluster analysis as a general topic and since it has been used in many fields of study, the novice user is faced with the daunting prospect of searching through a multitude of journals for appropriate references. In order to organize this diverse and voluminous material the following points will be considered: the terminology associated with cluster analysis; the journals containing the most significant papers; the broad nature of the published papers.

Finally, a few thoughts on the state of the literature of cluster analysis are given.

I. INTRODUCTION

Cluster analysis is unsupervised learning method that constitutes a corner stone of an intelligent data analysis process. It is useful for the exploration of inter-relationships among a collection of patterns, by organizing into homogeneous clusters. It is called as unsupervised learning because no a priori labeling of some patterns is available to use in categorizing others and inferring the cluster structure of the whole data. Intra-connectivity is a measure of the density. A high intra-connectivity means a good clustering arrangement because the instances grouped within the same cluster are highly dependent on each other. an Inter-connectivity is a measure of the connectivity between distinct clusters. A low degree of interconnectivity is advantageous because it indicates that individual clusters are largely independent of each other. Every instance in the data set can be represented using the same set of attributes. The attributes are categorical. To stimulate a hypothesis from a given data set, a learning system requires to make assumptions about the hypothesis to be learned. These assumptions are called as biases. Since every learning algorithm uses some biases, it reacts well in some domains where its biases are appropriate while it performs poorly in other domains. The problem with clustering methods is that the interpretation of the clusters may be difficult. The algorithms will always assign the data to clusters even if there were no clusters in the data. Cluster analysis is a difficult problem because many factors

1. Effective similarity measures,
2. Criterion functions,
3. Algorithms are come into play in devising a well tuned clustering technique for a given clustering problems.

Moreover, it is well known that no clustering method can adequately handle all sorts of cluster structures i.e shape, size and density. Sometimes the quality of the clusters that are found can be improved by preprocessing the given data. It is not uncommon to try to find noisy values and eliminate them by a preprocessing step. A common technique is to use post processing steps to try to fix up the clusters that have been found.

II. TERMINOLOGY

Terminology differs from one field to another. In biology, a significant field of study for the use of cluster analysis, the term **'numerical taxonomy'** is frequently used as a substitute for cluster analysis. In pattern recognition the terms are generally **'clustering'** or 'classification', but in cybernetics the term 'unsupervised learning' is often found.

Geographers use the term 'regionalization' and anthropologists sometimes use cluster analysis to solve a problem they call 'sedation'. Other terms which are frequently used in most fields are 'classification', 'grouping' and 'clumping', 'typology' and 'Q-analysis'.

Further differences in terminology occur throughout the formulation of problems and the description of algorithms, not so much as a result of the use of different words but usually as a result of the use of one word to mean different things. For example, classification is used by some authors to describe techniques for assigning individuals to groups having *a priori* labels and by others to describe the allocation of individuals to initially undefined groups. Fortunately, most authors seem to be acutely aware of this problem and state explicitly their intended meaning of such terms.

Table 1. Journals containing significant publications on cluster analysis

sno	**Title of the paper**	**Name of the journal**	**abstract**	**conclusion**
1	A Survey of Clustering Techniques for Big Data Analysis	IEEE-2014	In this paper discussion ofsome of the current big data mining clustering techniques was done. Comprehensive analysis of these techniques is carried out and appropriate clustering algorithm is provided	In this paper, there are various clustering Techniques which are currently used for analyzing big data. All these recent techniques are compared on the basis of execution time and cluster quality and their merits and demerits are provided.
2	Classification and Analysis of Clustering Algorithms for Large Datasets	IEEE-2015	This paper gives an overview of available algorithms that can be used for clustering in large datasets. The comparative analysis of available clustering algorithms is provided in this paper. This paper also includes the future directions for researchers in	This paper presents the basic classification of clustering algorithms. The comparison of k-means, single linkage, average linkage, complete linkage, BIRCH, DBSCAN and CLIQUE is given in this paper on the basis of some basic parameters. The available datasets that researchers can utilize to carry out the research

			the large database clustering domain.	in data mining and clustering domain are listed in this paper
3	Data Clustering Approaches Survey and Analysis	IEEE-2015	This paper is intended to examine and evaluate various data clustering algorithms. The two major categories of clustering approaches are partition and hierarchical clustering. The algorithms which are dealt here are:k-means clustering algorithm, hierarchical clustering algorithm, density based clustering algorithm, self-organizing map algorithm, and expectation maximization clustering algorithm. All the mentioned algorithms are explained and analyzed based on the factors like the size of the dataset, type of the data set, number of clusters created, quality, accuracy and performance. This paper also provides the information about the tools which are used to implement the clustering approaches. The purpose of discussing the various software/tools is to	After comparing the results based on all the factors mentioned, here are some of the conclusions which are observed: 1. The performance is inversely proportional to the number of clusters considered. 2. The performance of k-means and EM is better than Hierarchical and SOM algorithm. 3. SOM shows the highest accuracy in clustering the random data. 4. Hierarchical and SOM algorithms have better quality than the other two algorithms. 5. All the algorithms have ambiguity in some noisy data when clustered. 6. When using large dataset, Quality of k-means and EM becomes better and when using small dataset, SOM and hierarchical shows better results 7. Hierarchical and SOM gives better results

			make the beginners and new researchers to understand the working, which will help them to come up with new product and approaches for the improvement.	when random dataset is used. 8. K-means and EM algorithms are less resistant to noise. 9. All the algorithms give almost the same results when implemented in different software
4	An Analysis on Clustering Algorithms in Data Mining	IJCSMC-2014 Vol. 3, Issue. 1, January	This paper provides a broad survey of the most basic techniques and identifies .This paper also deals with the issues of clustering algorithm such as time complexity and accuracy to provide the better results based on various environments. The results are discussed on huge datasets	Given a data set, the ideal scenario would be to have a given set of criteria choose a proper clustering algorithm to apply. Choosing a clustering algorithm, however, can be a difficult task. Even finding just the most relevant approaches for a given data set is hard. Most of the algorithms generally assume some implicit structure in the data set. This paper provides a broad survey of the most basic techniques.
5	A Survey on Clustering Techniques for Big Data Mining	Indian Journal of Science and Technology(IJST) VOL-9 *January 2016*	This paper focuses on a keen study of different clustering algorithms highlighting the characteristics of big data. Brief overview of various clustering algorithms which are grouped under partitioning, hierarchical, density, grid based and model based are discussed.	This paper analyzed different clustering algorithms required for processing Big Data. The study revealed that to identify the outliers in large data sets, the algorithms that should be used are BIRCH, CLIQUE, and ORCLUS. To perform clustering, various algorithms can be used but to get appropriate results the present study suggests that – by using CURE and ROCK algorithms on categorical data, arbitrary shaped clusters will be

				created. By using COBWEB and CLASSIT algorithms on numerical data with model based, non-convex shape clusters can be formed. For spatial data STING, OPTIGRID, PROCLUS and ORCLUS algorithms when applied yield arbitrary shaped clusters.
6	A clustering approach using a combination of gravitational search algorithm and k-harmonic means and its application in text document clustering	Turkish Journal of Electrical Engineering & Computer Sciences **Accepted/Published Online:** 17.04.2016	In this paper, a combination method of GSA and K-harmonic means, called GSA-KHM, has been proposed, in which the dependency on the initialization has been improved. The proposed GSA-KHM method has been applied to data clustering. As a special application, it has also been used on the text document clustering application. The simulation results show that the proposed method works better than the GSA-KM and other comparative methods in both data clustering and text document clustering applications.	In this study, a clustering method has been presented by combination of the GSA and KHM. The proposed method has better clustering results than GSA-KM. In addition, unlike GSA-KM, this method is not dependent on the initial centers. This method was applied to five well-known UCI datasets and four textual datasets. The simulation results of both cases show better performance of the proposed GSA-KHM in comparison with GSA-KM
7	Comparative Study of Clustering Data Mining: Techniques and Research Challenges	IJLTEMAS Volume III, Issue IX, September 2014	In this paper covers the various clustering techniques. A tabular comparison of work done by various authors is presented. This paper reviews five types of clustering	Clustering is concern to cluster or categories the “similar” or “dissimilar” dataset into different groups. This paper focuses on the existing literature in the field of data mining clustering. From the analysis it was found that there is no

			data mining techniques- Partitioning Clustering, Hierarchical Clustering, Grid based clustering, Model based clustering, and Density based clustering	single technique is applicable/dependable in all domains. But still from analysis, we have conclude that K-means method perform better than other method in many domain.
8	A Survey of Data Mining Clustering Algorithms	International Journal of Computer Applications Volume 128 – No.1, October 2015	This paper aims to provide a brief overview and comparison of different clustering algorithms and methods. The different partitioning methods studied here are k-means and k-medoids. The different hierarchical techniques studied here are BIRCH and CHAMELEON. The different grid-based techniques which are described are DBSCAN and DENCLUE. Lastly, the different techniques which are used in grid-based technique, like STING and CLIQUE are described	This paper aims to provide an overview of the algorithms used in different clustering techniques along with their respective advantages and disadvantages. The different clustering methods that have been studied are partitioning clustering, hierarchical clustering, density based clustering and grid based clustering. 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
9	Study on Various Clustering Techniques	IJCSIT- Vol. 6 (3) 2015	The main aim of this review paper is to provide a comprehensive review of different clustering techniques in data	Clustering is that technique of data mining which is used to extract the useful information from raw data

			mining. Clustering is the subject of active research in many fields such as statistics, pattern recognition and machine learning. Cluster Analysis is an excellent data mining tool for a large and multivariate database.	
10	Customer Segmentation Using Clustering and Data Mining Techniques	International Journal of Computer Theory and Engineering, Vol. 5,No.6, December 2013	This research paper is a comprehensive report of k-means clustering technique and SPSS Tool to develop a real time and online system for a particular super market to predict sales in various annual seasonal cycles. The model developed was an intelligent tool which received inputs directly from sales data records and automatically updated segmentation statistics at the end of day's business. The model was successfully implemented and tested over a period of three months. A total of n= 2138, customer, were tested for observations which were then divided into k= 4 similar groups. The classification was based on nearest mean. An	In summary, the cluster analysis of the chosen sample of respondents explained a lot about the possible segments which existed in the target customer population. Once the number of clusters was identified, a k-means clustering algorithm, which is a non-hierarchical method, was used. For computing k-means clustering, the initial cluster centers were chosen and then final stable cluster centers were computed by continuing number of iterations until means had stopped further changing with next iterations. This convergent condition was also achieved by setting a threshold value for change in the mean. The final cluster centers contained the mean values for each variable in each cluster. Also, this was interpreted in multi-dimensional projections related to market forecasting and planning. To check the stability of the clusters, the sample

			ANOVA analysis was also carried out to test the stability of the clusters. The actual day to day sales statistics were compared with predicted statistics by the model. Results were quite encouraging and had shown high accuracy.	data was first split into two parts and was checked that whether similar stable and distinct clusters emerged from both the sub-samples. These analyses at the end provided further illustrations of using cluster method for market segmentation for forecasting.
11	A Survey on Different Clustering Algorithms in Data Mining Technique	International Journal of Modern Engineering Research Vol.3, Issue.1, Jan-Feb. 2013	Clustering is a kind of unsupervised data mining technique. It describes the general working behavior, the methodologies followed by these approaches and the parameters which affect the performance of these algorithms. In classifying web pages, the similarity between web pages is a very important feature. The main objective of this paper is to gather more core concepts and techniques in the large subset of cluster analysis	The cluster analysis examines unlabeled data, by either constructing a hierarchical structure, or forming a set of groups, according to a pre specified number. In this paper, an attempt has been made to give the basic concept of clustering, by first providing the definition of different clustering algorithms and some related terms
12	A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis	IEEE TRANSACTIONS ON EMERGING TOPICS IN OMPUTING June 2014	one of the major issues in using clustering algorithms for big data that causes confusion amongst practitioners is the lack of consensus in the definition of their properties as well as a lack of	This survey provided a comprehensive study of the clustering algorithms proposed in the literature In general, the empirical study allows us to draw the following conclusions for big data: •No clustering algorithm performs well for all the evaluation criteria, and

			formal categorization. With the intention of alleviating these problems, this paper introduces concepts and algorithms related to clustering, a concise survey of existing (clustering) algorithms as well as providing a comparison, both from a theoretical and an empirical perspective. From a theoretical perspective, we developed a categorizing framework based on the main properties pointed out in previous studies. Empirically, we conducted extensive experiments where we compared the most representative algorithm from each of the categories using a large number of real (big) data sets. The effectiveness of the candidate clustering algorithms is measured through a number of internal and external validity metrics, stability, runtime, and scalability tests.	future work should be dedicated to accordingly address the drawbacks of each clustering algorithm for handling big data. •EM and FCM clustering algorithms show excellent performance with respect to the quality of the clustering outputs, except for high-dimensional data. However, these algorithms suffer from high computational time requirements. Hence, a possible solution is to rely on programming language and advances hardware technology which may allow such algorithms to be executed more efficiently. •All clustering algorithms suffer from stability problem. To mitigate such an issue, ensemble clustering should be considered. •DENCLUE, OptiGrid and BIRCH are suitable clustering algorithms for dealing with large datasets, especially DENCLUE and OptiGrid, which can also deal with high dimensional data
13	Comparative Studies of Various Clustering Techniques and Its Characteristics	Int. J. Advanced Networking and Applications Volume: 5 Issue:6-2014	Discovering knowledge from the mass database is the main objective of the Data Mining. Clustering is the key technique in	Discovering knowledge from the mass database is the main objective of the Data Mining. Clustering is the key technique in data mining. A cluster is made up of a number

			data mining. A cluster is made up of a number of similar objects grouped together. The clustering is an unsupervised learning. There are many methods to form clusters. The four important methods of clustering are Partitional Clustering, Hierarchical Clustering, Density-Based Clustering and Grid-Based Clustering. In this paper, these four methods are discussed in detail	of similar objects grouped together. This paper gives the review of four important techniques namely Partitional Clustering, Hierarchical Clustering, Density Based Clustering and Grid Based Clustering. The different algorithms of these techniques are discussed. So this paper provides a quick review of these four clustering techniques.
14	Analysis and Application of Clustering Techniques in Data Mining	International Journal of Computing Algorithm Volume: 03, May 2014	The research about clustering makes a spurt development in recent years, and then produced a variety of clustering algorithm. WEKA is a data mining tool, it provides the facility to classify and cluster the data through machine learning algorithm. This paper analyses some typical methods of cluster analysis and represent the application of the cluster analysis in data mining	The overall goal of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters).Clustering can be done by the different no. of algorithms such as hierarchical, partitioning and grid algorithms.
15	Comparative Study of Clustering Techniques in Iris Data Sets		This paper analysis the four different data types clustering techniques like K-	The most efficient five types of data clustering techniques have been analyzed using iris flower data sets. The

			Means, Fuzzy c mean, Mountain clustering and Subtractive clustering in Iris flower data set. The accuracy, run time, time complexity are compared among them and then newly improved Y-means algorithm are proposed in order to improve the obtained clustering result using Matlab tool. The result shows that improved Y-means algorithm yields better result when compared to other clustering techniques with less computation time.	Estimation process are done for the accuracy of each algorithm and also the time complexity function are evaluated for performance measures. The improved Y-means has overcome the drawbacks of K-mean and other techniques

III. CONCLUSIONS

The literature of cluster analysis is scattered throughout many journals in many fields of study. The aim here was not to produce a comprehensive list of references, but to offer a guide to further reading in the subject and a selection of papers covering as many areas as possible. Numerous other references may be found in the papers cited at the end.

At present there is no journal devoted exclusively to the subject of cluster analysis. It is debatable whether or not there is a real need for one. Much can be learned by studying the articles published in the field of cluster analysis. The papers concerning the application of the techniques in particular, should give anyone contemplating using cluster analysis a good idea as to whether or not it will be a suitable and useful method of data analysis; and should help the theorists understand the practical problems involved in using these techniques.

REFERENCES

1. *MR. Anderberg, Cluster Analysis for Applications, Academic Press, London (1973).*
2. *T. W. Anderson, S. Das Gupta and G. P. H. Styan,» Bibliography of Multivariate Statistical Analysis, Oliver and Boyd, Edinburgh(1972).*
3. *H. T. Clifford and W. Stephenson, An Introduction to Numerical Classification, Academic Press, London (1975).*
4. *A. J. Cole, Numerical Taxonomy, Academic Press, London(1969).*
5. *B. S. Duran and P. L. Odell, Cluster Analysis. A Survey, Springer-Verlag, Berlin (1974).*
6. *B. Everitt, Cluster Analysis, Heinemann Education Books,London (1974).*
7. *W. D. Fisher, Clustering and Aggregation in Economics, JohnHopkins, Baltimore (1968).*
8. *J. A. Hartigan, Clustering Algorithms, John Wiley & Sons, New York (1975).*
9. *N. Jardine and R. Sibson, Mathematical Taxonomy, John Wiley & Sons, London (1971).*
10. *R. R. Sokal and P. H. A. SneatVi,Numerical Taxonomy, Freeman,San Francisco (1973).*

11. *R. C. Tryon and D. E. Bailey, Cluster Analysis, McGraw-Hill, New York (1970). (This book is now out of print.)*
12. *J. Van Ryzin, Classification and Clustering, Academic Press, New York (1977*