One Day National Conference on"Recent Innovations in Science, Technology and Engineering"
National Institute of Technology (NIT) , Srinagar, Jammu and Kashmir (India) RISTE-2017
16th December 2017, www.conferenceworld.in ISBN: 978-93-86171-87-0

Clustering Techniques in Data Mining : A Survey

Naveen Trivedi[1], Aparna Shukla[2], Suvendu kanungo[3]

[1, 2, 3]*Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi, Allahabad Campus, (India)*

ABSTRACT

It becomes a very challenging task to find the interesting patterns from the high dimensional data growing drastically due to insufficient resolution, image corruption, and noisy data etc. Various data mining techniques are developed to find the relevant group and sample in the underlying data often consist of thousands of measurements. Clustering is one of the most popular techniques used to partitioning the data and extracting relevant patterns on the basis of some precise feature. It is one of the technique that meets today's era challenges efficiently. This paper is a survey paper briefing on clustering algorithms and applications. Well-known clustering techniques are discussed in this paper. The authors present the fundamental of data mining, clustering concepts, its algorithm and mention some applications of clustering. Moreover, the types of hierarchical algorithm are elaborated. Additionally, the variations of partitioning algorithm, k-means and k-medoids are also described. The authors attempt to cover the various aspects of clustering to provide a comprehensive knowledge of clustering techniques.

Keywords—clustering, clustering techniques, clustering types, data mining

I.INTRODUCTION

Nowadays to the development of an advanced technology of data collection, the large amount of data is collected from various sources and the analysis of this large amount of data are measured in billions every day, The traditional analysis approaches are not suitable for such a vast amount data and inhibit us to apply these traditional techniques. To extract the useful information on the datasets and grouping the important data become demanding day by day. With the high-dimensional amount of data resides in files, databases, and other repositories, it necessitates to develop robust techniques for inspecting, interpreting data and for the extraction of relevant knowledge that assist in making the decision.

1.1. Data Mining and Its Taxonomy

Data mining is the process to extract the hidden previously unidentified relevant patterns such as unusual records (anomaly detection), cluster analysis and dependencies [1] [2]. Relying on the background and view of the definers, some definitions of the data mining mentioned in the literature are discussed below:

Despite lofty definitions, Garten defines- Data mining as a process of discovering essential correlations, patterns, and trends by moving through a bulk of data stored in repositories [3]. At the same time, David Bolton gave the new definition of data mining. According to him, it is the process of processing voluminous data stored in the database, seeking for patterns and affiliation within that data [2]. SAS defines data mining as the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns of a business advantage [4]. The diverse operation in data mining is shown below in Fig.1.

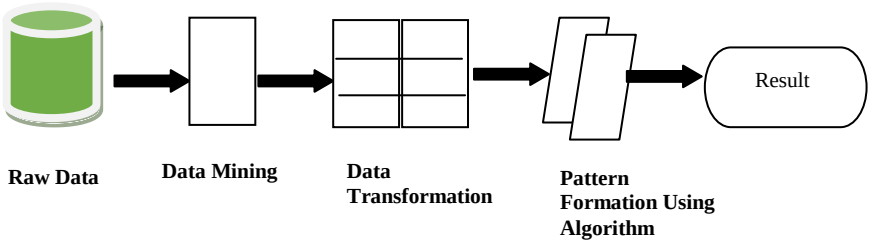

Figure 1. Operations in Data Mining

1.2. Data Mining and KDD

Identification and extraction of relevant hidden patterns and information from a bulk of data have been given a variety of names, including Data Mining (DM), text mining, Information Retrieval (IR), extraction of knowledge, Information harvesting, data archaeology and data processing.

Data mining term has been frequently used by statisticians, data analysts and other informatics communities. It extensively gained popularity in the database field too. In general, Data mining is used as a synonym for Knowledge discovery in database (KDD). The phrase Knowledge discovery in database (KDD) was first coined in 1989 [5]. DM also widely known as one of the crucial steps in KDD process, adduce (refer) to the nontrivial discovering of implicit, formerly concealed (unidentified) and potentially relevant information from the abundant data exceeding day by day. It is prevalent to discover the patterns from the abundant data. The following Fig. 2 below shows data mining as a core step in an iterative knowledge discovery process [5].

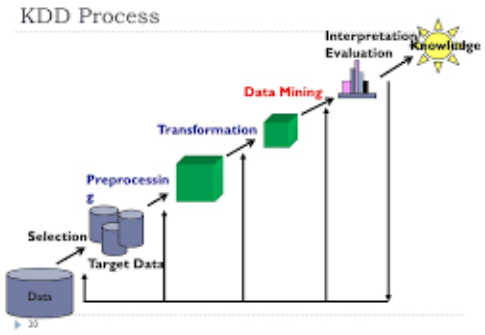

Figure 2. Data Mining as a KDD Process

1.3. Clustering

Clustering is most suitable techniques to distribute the data into groups of similar objects which are closely related to each other and different with the objects in the other groups. The clustering approaches smoothly arrange a set of patterns into the group or clusters on the basis of similarity measures. Cluster techniques are based on an unsupervised approach where data items are unlabeled to group them into valid clusters [1] [2], while in unsupervised approaches, the dataset is given in the form of pre-classified item set. If the dataset is already labeled it help us to create a new label.

This paper is a survey paper organized in different following sections: Second section presents a fundamental concept, various stages and some defined cluster definitions. The Third section of this survey paper emphasized on different types of clustering algorithms followed by its applications in various research domains in section IV. Finally, the conclusion statement of the paper is in the last section.

II.CLUSTERING AND ITS CONCEPT

This section of the paper provides a comprehensive review of the clustering and its notation. Clustering is one of the popular unsupervised data mining technique that partition the volume of data into a k clusters based on some homogeneity criterion. Several definitions have been discussed in the literature of clustering. According to Everett [1974], a cluster is a set of entities which are identical in some manner and entities from different clusters are not identical. The clustering process has to discover the latent structure in the data. It is one of the crucial steps in data mining process to segregate a data sets or objects into a set of significant sub-classes, called clusters [7]. Therefore, clusters described as a collection of objects which are similar between them and dissimilar to the set of objects belong to another cluster [8]. Consequently, a good cluster can be examined by having minimum intra-cluster distance and maximum inter-cluster distance.

It is the main function of the exploratory data mining which is used in diverse fields, including image segmentation, pattern analysis, text mining, machine learning techniques, bioinformatics and much more. Many of the research communities used the term "clustering" as to describe methods for aggregation of unlabeled data [2]. Different communities gave a different taxonomy and made assumptions regarding components of clustering process and the context in which the clustering is used. Therefore, mostly confront a dilemma regarding the scope of its survey.

Clustering is most ordinarily machine learning technique used to discover patterns from the sample data and placing data elements into analogous groups without any prior knowledge of the group definitions. Moreover, it can also be referred as data segmentation in some of the applications since it partitions large data sets into groups according to their resemblance. Its significance increases while detecting an outlier.

2.1 Clustering Stages

The various stages in clustering algorithm are shown in Fig. 3. Below:

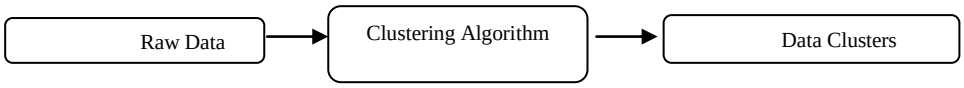

Figure 3. Clustering Stages

2.2 Some working definition of a Cluster

The Cluster does not have some peculiar definitions. However, some working definitions of clusters are mentioned below:

2.2.1 Well-Separated Cluster Definition

A cluster is said to be well-separated cluster so that each object is significantly closer to every other object in the cluster as compared to any other object that not lie in the cluster.

2.2.2 Center- based Cluster Definition

A cluster is a group of objects so that an object in a cluster is significantly nearest to the "center" of a cluster than to the center of any other cluster. The center of a cluster is often referred as the centroid.

2.2.3 Contiguous Cluster Definition

A cluster is a group of objects so that an object in a cluster is more similar to one or more other objects in the cluster as compared to any other object that does not belong to the same cluster. It also refers as *Transitive clustering.*

2.2.4 Density-based Cluster Definition

The two highly dense regions of cluster objects are separated by low-density regions. This is more applicable when noise and outliers are present and also when the clusters are irregular.

2.2.5 Similarity-based Cluster Definition

Grouping of clusters that share some common property or represent a particular concept. In other words, the objects belong to a single group that exhibits similarity in some fashion whereas the objects that are not similar belongs to another group.

III.CLUSTERING CLASSIFICATION

Several clustering techniques are discussed in the literature [1] [7]. Each of these algorithms may produce a different aggregation of a dataset. The choice of clustering is a crucial decision that relies upon several factors includes the type of output desired, the performance of method with types of data used, size of the dataset and the system specification are to be used [9]. Broadly the clustering techniques are classified into following categories as shown in Fig. 4. We primarily focus on hierarchical and partitioning clustering techniques.

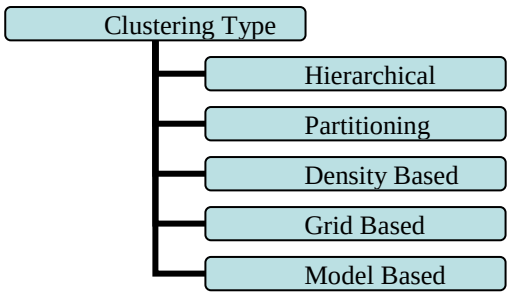

Figure 4. Clustering Classification

3.1 Hierarchical Clustering

Hierarchical clustering can be done either the top down or bottom up. On the basis of a way of processing, the hierarchical clustering is further divided into two categories as shown in Fig. 5. Agglomerative (bottom-up) and divisive (top-down).

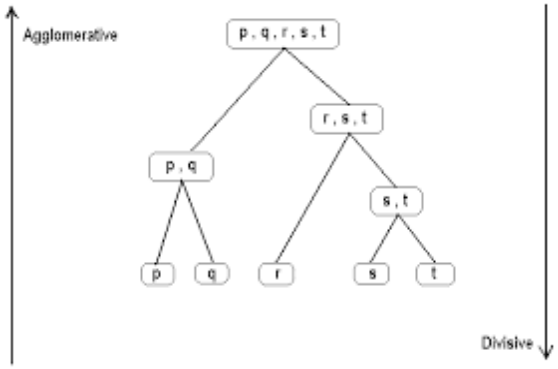

Figure 5. Hierarchical Clustering Types

Hierarchical clustering methods build a set of nested clusters in which each pair of clusters is progressively enclosed into large clusters until achieving a single cluster. This type of clustering techniques builds a tree of clusters, known as the dendogram. Every cluster node contains the child nodes and sibling clusters that concede data exploring at different levels of granularity.

3.1.1 Agglomerative Hierarchical Clustering (AGNES)

An agglomerative hierarchical clustering begins with the single point (singleton) and recursively by applying fusion operation combines two or most appropriate points, where a point is either an individual or a cluster itself [9]. Informally an agglomerative hierarchical clustering algorithm may be elaborated as follows:

Algorithm 1:

Let consider X be the set of n objects.

1. Begin with the disjoint clustering having level L (0) = 0 and sequence number h = 0.
2. Compute dissimilarity matrix for all inter-object
3. Find the closest pair of clusters; say pair (a), (b), depending on the homogeneity criterion.
4. Increment: h = h +1.
5. Apply fusion operation to merge clusters (a) and (b) into a single cluster to form the next clustering and setting the level of this clustering too.
6. Re-determine dissimilarities between new cluster and other objects or clusters.
7. Repeat step 3 until a single cluster of all objects is achieved.

Several Agglomerative techniques discussed in past so far due to disparate ways of characterizing similarity between clusters. The three different agglomerative techniques are a single linkage, complete linkage, and

average linkage. Single linkage agglomerative hierarchical clustering methods are the simplest one, also refer as nearest neighbor technique.

3.1.2 Divisive Hierarchical Clustering (DIANA)

A divisive hierarchical clustering is the reverse process of AGNES. It starts with the one cluster having all data points and recursively splits into clusters until k clusters are obtained [10]. The merging or division process is done on the basis of some similarity measures.

Advantages

- Provide easiness to handle any forms of similarity.
- Consequently, suitable to the diverse type of attribute.
- Embedded flexibility regarding the level of granularity.

Disadvantages

- Comprehension of the hierarchy is complex and often confounding.
- High complexity
- No objective function is minimized directly.
- It is difficult to find the number of clusters by the use of dendogram.

3.2 Non- Hierarchical Clustering (Partitioning Clustering)

The ground of non-hierarchical clustering is to partition the whole dataset into k predefine portions.

Given a dataset $D = \{d_1, d_2, ..., d_n\}$ consist of n objects and predetermined K such that $K \leq n$, the number of clusters to create. The partitioning algorithm partitions the dataset D into K clusters with the target to optimize the objective partitioning criterion such as to minimize the intra cluster distance and maximize the inter cluster distance.

These algorithms are iterative algorithm that iteratively reallocate in order to improve the partition by transiting the objects from one cluster to another [12]. Partitioning clustering techniques initiates with an initial partition then endeavor all probable shifting of objects from one class to another in order to optimize the objective functions.

The key factor of this type of clustering algorithms is creating groups either by the centroid of the clusters such as K-means, or by one object located centrally such as K-medoids. Therefore, also refer as the centroid-based clustering algorithms that partition the search space by finding the centroid of the clusters and recomputed it until convergence criterion met. These type of techniques are useful when there is a predefined number of groups. A problem with these algorithms is that they suffer from the combinatorial explosion due to the number of possible solutions i.e. it is computationally infeasible to test each possible subset of solutions. Specifically, leads towards the concept of relocation schemes that iteratively reallocate objects between the clusters K. Relocation algorithms progressively improve the clusters quality, unlike to hierarchical clustering methods.

The different approaches are taken into account for partitioning the dataset into K clusters. By the conceptual point of view, one approach of partitioning is to determine the clusters with a certain model whose unidentified

parameters have to be found such as *probabilistic clustering*. Another approach with the concept of *objective function* depending on a partition.

Iterative optimization partitioning algorithms are appropriately sub-divided into K-means and K-medoids methods. The general procedure of iterative relocation partitioning algorithm is mentioned in Algorithm 2.

Algorithm 2- I*terative Relocation Partitioning Algorithm*

Input:

$D = \{d_1, d_2, ..., d_n\}$// database contains n objects

K // the number of clusters, and $K \leq n$

Output:

A set of K clusters that minimizes the partition criterion function

Steps:

1. Initiates with the first K centers as the initial solution.
2. Compute the association for the data points using the current cluster centers.
3. Update some or all cluster centers according to new association of the data points.
4. Repeat Step 2. Until no change to partition criterion function or no transition of data points cluster.

There are several partitioning based clustering techniques discussed in literature. Most commonly used partitioning techniques are K-means [13], Fuzzy C-means, K-medoids, CLARA, DBSCAN and many other [8]. The K-means and K- medoids methods are elaborated further below in this paper.

3.2.1 K-means algorithm

K-means is the most popular clustering technique widely accepted by researchers to partition the dataset into K predefined clusters. It is used in scientific and industrial applications widely.

The ground idea is to classify the given set of data in K predefined clusters with the objective to optimize. This algorithm executes in two phases: *Initialization Phase* and *Assignment Phase*. In initial phase, the K objects are chosen randomly as the initial centroid. Further, in latter phase the objects are assigned to appropriate centroids to form well separated clusters some homogeneity measures criterion to minimize intra- cluster distance and maximize inter-cluster distance. Formally, K-means algorithm is described by algorithm 3.

Algorithm 3- K-means *Algorithm*

1. Select randomly K seed points as the K-initial centroid.
2. Compute dissimilarity matrix for K- initial centroid to all data objects of dataset given using some distance metric.
3. Assign objects to the closest centroid clusters.
4. Aster assigning all data objects to its nearest clusters, re-compute new centroid of each clusters.
5. Repeat step 2 until convergence criterion met.

This algorithm suffers mainly from the two potential drawbacks are: (1) randomly selection of K centroids that may leads to different results at different iteration , and (2) the prior knowledge of significant K must be made known beforehand.

3.2.2 K-medoids algorithm

K-means algorithm is failed to find outliers effectively because the mean values are easily influenced. Thus, the variant of K-means, K-medoids is introduced which handle the problem of noise and detection of outliers easily and efficiently. Rather than taking the mean of the value to represent the cluster, the K-medoids algorithm uses an actual value in the cluster. Medoids refers to the most centrally located objects in the cluster in such a way to minimize the sum of distances to other data points. K-medoids produce correctly represented cluster center and robust outliers in contrast to the K-means that cannot form correct cluster center. K-medoids method is similar to the K-means algorithm except when filling the cluster centers. K-medoids algorithm is useful in those applications where it is needed to be each center be the one of its cluster itself. Formally the K-medoids algorithm is described by algorithm 4.

Algorithm 4- K-medoids Algorithm

1. Select randomly K data points as the K-initial medoids.
2. Repeat
   i. Assign each remaining data objects to the nearest medoid.
   ii. Randomly select the non-medoids data and compute the total cost of swapping old medoids data objects with the currently chosen non-medoids data objects.
   iii. If total cost (after swapping) < 0, perform the swap operation to generate the set of new K-medoids
3. Stop when medoids stabilize their locations.

Several approaches are existing in past so far to perform K-medoids clustering such as PAM, CLARA, CLARANS.

3.3 Density Based Clustering

The efficient computability in the context of high dimensional data sets has come into focus and led to algorithm variants of computing the clustering model based on density.

In density-based clustering methods, a cluster is a set of data objects spread over a high density of objects, separated from the other density-based clusters spread over the region of the low density of objects. These separated areas of clusters usually represent the outliers or the noise. A point p from point q can be either density reachable or density connective.

Density Reachability:

A point "p" is said to be density-reachable from density"q", if the point "p" lies within the ε distance from the point"q" and "q" has a sufficient number of points as its neighbour which are within distance ε .

Density Connectivity:

There exist a point "r" such that it has a sufficient number of points as its neighbours and also both the point "r" and "q" lies within the distance ε, then a point "p"is said to be density connected with the point"q". This is so-called the chaining process, if the"q" is neighbour of"r", "r" is neighbour of "s" which in turn neighbour of"p", implies that "q" is neighbour of"p".

These approaches are nonparametric approach and thus neither requires input parameter K, the number of clusters nor make any assumptions concerning the underlying density say. Apart from these advantages of density-based clustering algorithm such as DBSCAN, OPTICS [14], and these algorithms also discover arbitrarily shape of clusters. But these type clustering model are not suitable for neck type of dataset.

3.4 Grid Based Clustering

Grid-based clustering algorithms quantize an object space into the set of a fixed number of cells to form a multi-resolution grid structure in which the objects are allotted to the suitable grid cell [16]. Further, the density of each cell is computed and the cell whose density lies below a predefined threshold value t is eliminated.

The principal advantage of grid-based clustering approaches such as STING and CLIUE is its fast processing time, which is typically dependent on the number of cells populated in the quantized space and independent the number of data objects in the dataset. Hierarchical type of cluster structure is developed while performing the grid-based clustering and performs efficiently to reply various queries.

3.5 Model Based Clustering

The traditional clustering such as partitioning and hierarchical are heuristic and does not base on any formal models. Model-based clustering methods are alternative methods based on some formal models. These clustering techniques consider that the data is drawing from a distribution that is the mixture of two or more clusters. This clustering uses the soft assignment, which refers that each data object has a probability of membership to each cluster [8] [11].

This method yields robust clustering technique since it offers several potential advantages over other heuristic clustering techniques are:

- Facilitate an efficient way to identify the number of clusters automatically based on standard statistics,
- Handle the problem of outlier or noise considerably.

IV.CLUSTERING APPLICATION

Clustering has several applications widely in many research domain such as text mining, image analysis, economic, science, pattern recognition, data mining and so many other. Some of the most important applications of the clustering are discussed in this section

4.1 Clustering in data mining

It is one of the crucial steps in data mining process. It discovers and recognized the related groups that serve as the initial point and used for further relationships identification.For example: help marketers to identify the

distinct groups in their customer bases and further use this information to develop their market strategy accordingly in order to achieve maximum profit.

4.2 Clustering in text mining

Text mining also referred as Text data mining is a process to retrieve quality information form the text. This mining usually follows a series of steps includes the inputting text, deriving a pattern from the inputted text after parsing the structured data, and finally performing evaluation and representation of the result in a systematic manner.

Typically, text mining tasks involve the phases: text categorization, text clustering, concept extraction, summarization and lastly modeling entity relation. For example, search navigation system, in which the query is imposed into the system and in response extract the relevant document information related to the query given.

4.3 Some other applications

Cluster analysis is done in the banking system to refine the user group, in industry, in insurance according to the type of residence and other features. In insurance company to identify the group of specific type of insurance policyholders with a high average claim cost. In the internet, cluster analysis is done for document classification and information retrieval [6].

In biotechnology, cluster analysis is done to categorize animal and plant according to population and in order to obtain the latent structure of knowledge. In geography, the biologists make use of clustering to find the species and their relationship. Clustering also plays a vital role in biometric identification system by reducing search space.

V.SUMMARY

The main objective of data mining process to retrieve the latent relevant pattern from the voluminous data and represent it in presentable and understandable form. Clustering is one of the important data mining approaches among the existing one which discover the cluster by partitioning the whole data set into k predefined clusters with the optimization objective function. This survey paper attempts to focus on clustering and its basic fundamental, along with the classification of clustering methods. The last section of the paper also attempted to highlight the different application domain of clustering widely used.

REFERENCES

[1] P. Berkhin, " A Survey of Clustering Data Mining Techniques", In: Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data. Springer, Berlin, Heidelberg, pp.25-71, 2006.

[2] E. Hruschka, R. Campello, A. Freitas, A. Carvalho, "A Survey of Evolutionary Algorithm for Clustering", IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 39, no. 2, pp. 133-155, 2009.

[3] M. Filippone, F. Camastra, F. Masulli, S. Rovetta, " A survey of kernel and spectral methods for clustering", Pattern Recognition, vol. 41, no. 1, pp. 176-190, 2008.

[4] J. Borland, J. Hirschberg, J. Lye, "Data Reduction of Discrete Responses: An Application of Cluster Analysis", Applied Economics Letters, Taylor and Francis Journals, vol. 8, no. 3, pp. 149-53, 2001.

[5] U. Fayyad, G. Piatetsky- Shapiro, P. Smyth, "The KDD process for extracting useful knowledge from volumes of data", Communications of the ACM, vol. 39, no. 11, pp. 27–34, 1996

[6] M. R. Anderberg, " Cluster analysis for applications" New York : Academic Press, 1973.

[7] J. A. Hartigan, "Direct clustering of a data matrix", Journal of the American Statistical Association, vol. 67, no. 337, pp. 123–129, 1972.

[8] P. Rai, S Singh, "A Survey of Clustering Techniques", International Journal of Computer Applications, vol. 7, no. 12, pp. 1-5, 2010.

[9] Dr L. Arockiam, S.S.Baskar, L. Jeyasimman, "Clustering Techniques in Data mining", Asian Journal of information Technology, vol. 11, no. 1, pp. 40-44, 2012.

[10] A. K. Jain, R. C. Dubes, "Algorithms for Clustering Data", Prentice-Hall, Englewood Cliffs, NJ, 1988.

[11] L. Kaufman, P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis", Wiley Interscience: John Wiley and Sons, New York, NY, 1990.

[12] R. Agrawal, J. Gehrke, D. Gunopulos,P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications", International Conference on Management of data (ACM Proceeding SIGMOD'98), vo. 27, no. 2, pp.94-105, 1998.

[13] R. Sharma, M. A. Alam, A. Rani , "K-Means Clustering in Spatial Data Mining using Weka Interface" , IJCA Proceeding on International Conference on Advances in Communication and Computing Technologies (ICACACT), vol. 1, pp. 26-30, 2012

[14] P. Shrivastava, H. Gupta. "A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (ISSN (print), vol.2, no. 3, pp. 200-202, 2012.

[15] M.Vijayalakshmi, and M. R. Devi, " A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets" , International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 3, pp.305-307, 2012.

[16] W. Liao, Y. Liu, A. Choudhary, "A Grid-based Clustering Algorithm using Adaptive Mesh Refinement", Appears in the 7th Workshop on Mining Scientific and EngineeringDatasets, pp.1-9, 2004.

[17] D. Sisodia, L. Singh, S. Sisodia, K. Saxena, "Clustering Techniques: A Brief Survey of different clustering Algorithms", International Journal of Latest Trends in Engineering and Technology (IJLTET), vol. 1, no. 3, pp. 82-87 , 2012.

[18] K. Wong, "A Short Survey on Data Clustering Algorithms", Second International Conference on Soft Computing and Machine Intelligence, IEEE, 2015.

T. Sajana, C. M. Sheela Rani and K. V. Narayana, "A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology,vol. 9, no.3, 2016.