International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600

http://www.ijcsjournal.com Volume 5, Issue 1, No 19, 2017 ISSN: 2348-6600
Reference ID: IJCS-247 PAGE NO: 1560-1571

Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

A REVIEW ON CLUSTERING TECHNIQUES

C.Solaiyappan1*,
MCA student ,JJ College of arts & science
Pudukkottai
E-Mail: solaiyappanc1995@gmail.com

L.Prisilla2*
Assistant Professor
JJ College of arts & science
Pudukkottai
E-Mail: presirose@gmail.com

Abstract - Data mining is the breakthrough of identifying the hidden patterns. Data mining is sorting through data to finding the patterns, anomalies and correlations within large data sets to predict outcomes. Clustering is the main task of exploratory data analysis and data mining applications. Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions and it is the process of making a group of abstract objects into classes of similar objects.clusters has the creditability to pinpointing the natural groupings of cases based on a set of attributes clustering is used in various application in the real world, Such as data/text mining, voice mining,Image processing, web mining.Clustering can be done by various of algorithms such as hierarchical, partitioning, grid and density based algorithms. In this paper, a review of diversified clustering techniques in data mining is presented.

KEYWORDS— Clustering, Types of Clustering.

1.INTRODUCTION

The process of digging the data to discover hidden connections and predict future trends has a long history . It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The construction of a data warehouse, which involves data cleaning and data integration, can be viewed as an important pre-processing step for data mining. Data mining uses the data warehouse as the source of information for knowledge data discovery (KDD) systems. The major steps involved in a data mining process are:

- Extract, transform and load data into a data warehouse

- Store and manage data in multidimensional databases

- Provide data access to business analysts using application software

- Present analyzed data in easily understandable forms, such as graphs

Data mining has following vital tasks Association, Sequence or path analysis, Classification, Clustering, Forecasting, Regression, Anomaly detection.

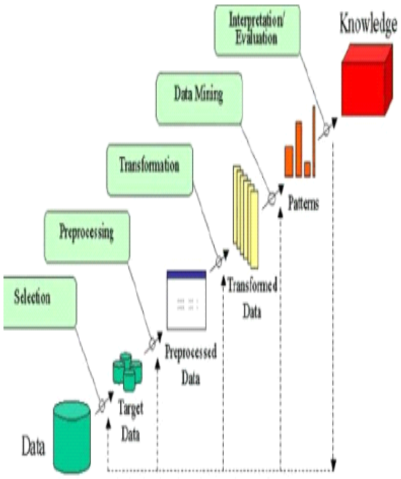

Figure 1

2. CLUSTERING

Clustering is an important technique in data mining and it is the process of partitioning data into a set of clusters such that each object in a cluster is similar to another object in the same cluster, and dissimilar to every object not in the same cluster. Clustering is often one of the first steps in data mining analysis Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures. Clustering analyses the data objects without consulting a known class label. This is because class labels are not known in the first place, and clustering is used to find those labels. Good clustering exhibits high intra-class similarity and low inter-class similarity, that is, the higher the similarity of objects in a given cluster, the better the clustering. The superiority of a clustering algorithm depends equally on the similarity measure used by the method and its implementation. The superiority also depends on the algorithm's ability to find out some or all of the hidden patterns . The different ways in which clustering methods can be compared are partitioning criteria, separation of clusters, similarity measures and clustering space

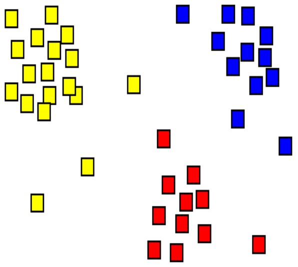

Figure 2

In data mining the data is mined using two learning approaches

i.e.

1) supervised learning

2)unsupervised clustering.

In this training data includes both the input and the desired results. These methods are fast and accurate. The correct results are known and are given in inputs to the model during the learning process. Supervised models are the neural network, Multilayer Perceptron, Decision trees.

Unsupervised Learning

The model is not provided with the correct results during the training. It can be used to cluster the input data in classes on the basis of their statistical properties only. unsupervised models are different types of clustering, distances and normalization, k-means, self organizing maps.

From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others.

3. CLUSTERING ALGORITHMS

Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms[1]. These methods vary in (i) the procedures used for measuring the similarity (within and between clusters) (ii) the use of thresholds in constructing clusters (iii) the manner of clustering, that is, whether they allow objects to belong to strictly to one cluster or can belong to more clusters in different degrees and the structure of the algorithm .

The different clustering techniques are stated as follows:

- Partitioning clustering
- K- Means
- K-Medoids
- PAM
- CLARA
- Hierarchical clustering
- Agglomerative
- BIRCH
- CHAMELEON
- Divisive
- Density-based clustering
- DBSCAN

- DENCLUE
- OPTICS
- Grid-based clustering
- STING
- CLIQUE
- Graph based clustering

3.1.PARTITIONING CLUSTERING

In the partitioning algorithm, the data is split into k partitions, where each partition represents a cluster and k<=n, where n is the number of data points. Partitioning methods are based on the idea that a cluster can be represented by a centre point.

Two conditions that must be satisfied in partitioning algorithms are: (i) At least one object should be present in one group or cluster and (ii) Each object must belong to only one cluster .

Partitioning algorithms divide data into several subsets. The reason of dividing the data into several subsets is that checking all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes are used in the form of iterative optimization. Specifically, this means different relocation schemes .

That iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters. There are many methods of partitioning clustering; they are k-mean,Bisecting K Means Method, MedoidsMethod, PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications) and the Probabilistic Clustering.

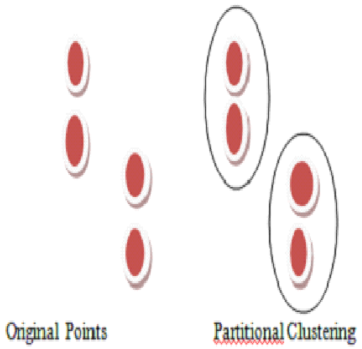

Figure 3

3.1.1.K-MEANS

In k-means algorithm, a cluster is represented by its centroid, which is a mean(average pt.) of points within a cluster. This works efficiently only with numerical attributes. And it can be negatively affected by a single outlier. The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis which aims to partition „n‟ observations into k clusters in which each observation belongs to the cluster with the nearest mean.The time complexity of k-means is O (nkt) , where n is the total number of objects, k is the number of clusters, and t is the number of iterations . The clusters formed by k-means have convex shapes.

The basic algorithm for k-means clustering is as follows :

The basic algorithm is very simple

1. Select K points as initial centroids.

2. Repeat.

3. Form K clusters by assigning each point to its closest centriod.

4. Recompute the centroid of each cluster until centroid does not change.

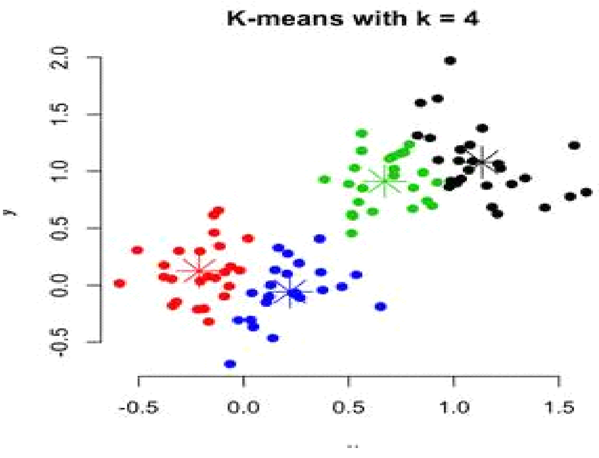

Figure 4

The k-means algorithm has the following important properties:

1. It is efficient in processing large data sets.

2. It often terminates at a local optimum.

3. It works only on numeric values.

4. The clusters have convex shapes.

DISADVANTAGES OF K-MEANS

4) Usually terminates at the local optimum, and not the global optimum. Can only be used when the mean is defined and hence requires specifying k, the number

3.1.2.K-MEDOIDS

In this algorithm we use the actual object to represent the cluster, using one representative object per cluster. Clusters are generated by points which are close to respective methods. The partitioning is done based on minimizing the sum of the dissimilarities between each object and its cluster representative .

PAM

Like all partitioning methods, PAM works in an iterative, greedy way. The initial representative objects are chosen randomly, and it is considered whether replacing the representative objects by non-representative objects would improve the quality of clustering. This replacing of representative objects with other objects continues until the quality cannot be improved further. PAM searches for the best k-medoids among a given data set. The time complexity of PAM is O(k(n-k)2) [3]. For large values of n and k, this computation becomes even more costly than the k-means method.

ALGORITHM

a) Arbitrarily choose k objects in D as the initial representative objects or seeds.

b) Repeat

i) Assign each remaining object to the cluster with the nearest representative object

ii) Randomly select a non-representative object, random

iii) Compute the total cost, S, of swapping representative object with a random

iv) If S<0 then swap on with random to form the new set of k representative objects.

c) Until no change

CLARA

CLARA uses 5 samples, each with 40+2k points, each of which are then subjected to PAM, which computes the best medoids from the sample [5]. A large sample usually works well when all the objects have equal probability of getting selected. The complexity of computing medoids from a random sample is O(ks2+k(n-k)), where s is the size of the sample . CLARA cannot find a good clustering if any of the best sampled medoids is far from the best k-medoids.

3.2 HIERARCHICAL CLUSTERING

Hierarchical clustering builds a cluster hierarchy (or a tree of clusters), called as dentogram. This method is based on the connectivity approach based clustering algorithms. In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- Agglomerative: This is a “bottom up” approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

- Divisive: This is a “top down” approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

3.2.1. AGGLOMERATIVE CLUSTERING

It is also known as AGNES. An agglomerative cluster starts with singleton clusters and recursively merges two or more most appropriate clusters. The algorithm forms clusters in a bottom-up manner, as follows :

a) Initially, put each article in its own cluster.

b) Among all current clusters, pick the two clusters with the smallest distance.

c) Replace these two clusters with a new cluster, formed by the smallest distance.

d) Repeat the above two steps until there is only one remaining cluster in the pool.

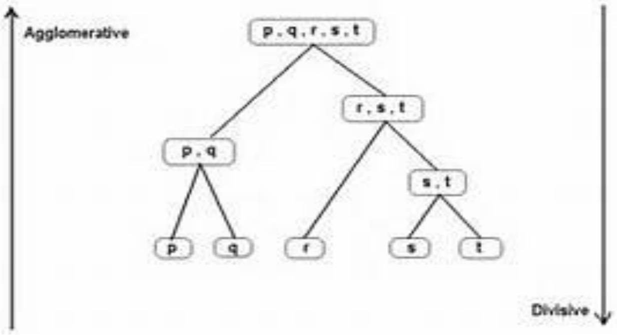

Figure 5

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):

BIRCH is an agglomerative hierarchical based clustering algorithm. It is used for clustering large amounts of data. It is based on the notion of a clustering feature (CF) and a CF tree. A CF tree is a height-balanced tree. Leaf nodes consist of a sequence of clustering features, where each clustering feature represents points that have already been scanned. It is mainly used when a small number of I/O operations are needed. BIRCH uses a multi clustering technique, wherein a basic and good clustering is produced as a result of the first scan, and additional scans can be used to further improve the quality of clustering . The time complexity of BIRCH is O(n) where n is number of clusters .

CHAMELEON

CHAMELEON is an agglomerative hierarchical clustering technique where, unlike other algorithms which use static model of clustering, CHAMELEON uses dynamic modeling to merge the clusters. It does not need users to provide information, but adjusts the clusters to be merged automatically according to their intrinsic properties. It has complexity of O(Nm+ N log (N)+m2log(m)), where N is the number of data points and m is the number of partitions . CHAMELEON uses the interconnectivity and compactness of the clusters to find out the similarity between them. It can be used to find clusters of varying densities . However, the processing time for high-dimensional data may go up to O(n2).

3.2.2DIVISIVE CLUSTERING

It is also known as DIANA. A divisive cluster starts with one cluster of all data points and recursively splits the most appropriate cluster. This process continues until a stopping criterion (usually the number of requested clusters k) is reached . It uses a top-down strategy. The algorithm for divisive clustering is as follows :

- Put all objects in one cluster.
- Repeat until all clusters are singletons
- Choose a cluster to split
- Replace the chosen cluster with the sub-cluster.

Advantages of hierarchical clustering

- Embedded flexibility regarding the level of granularity.

- Ease of handling any forms of similarity or distance.

- Applicability to any attributes type.

Disadvantages of hierarchical clustering

- Vagueness of termination criteria.

- Most hierarchal algorithm do not revisit once constructed clusters with the purpose of improvement.

4.DENSITY BASED CLUSTERING

In density-based clustering, clusters are defined as areas of higher density than the remaining of the data set. Objects in these sparse areas – that are required to separate clusters – are usually considered to be noise and border points. There are two major approaches for densitybased methods. The first approach pins density to a training data point and is reviewed in the sub-section Density-Based Connectivity. In this clustering technique density and connectivity both measured in terms of local distribution of nearest neighbors. So defined density-connectivity is a symmetric relation and all the points reachable from core objects can be factorized into maximal connected components serving as clusters. Representativealgorithms include DBSCAN, GDBSCAN, OPTICS, andDBCLASD.

The second approach pins density to a point in the attribute space and is explained in the sub-section Density Functions. In this, density function is used to compute the density.

Overall density is modeled as the sum of the density functions of all objects. Clusters are determined by density attractors, where density attractors are local maxima of the overall density function. The It includes the algorithm DENCLUE

DBSCAN (Density-Based Spatial Clustering of Application with Noise)

This algorithm is based on the user defined parameters, and on the same database with different parameters, it can create multiple clusters. The number of clusters is not required initially, because it produces the clusters only on the density basis. The data points in DBSCAN fall into three categories:

- Core points i.e. points that are at the interior of a cluster,

- Boundary points i.e. non-core points inside a boundary and

Outliers i.e. points that are neither core nor boundary points. DBSCAN cannot handle clusters of different densities .The basic idea of DBSCAN algorithm is that a neighborhood around a point of a given radius must contain at least minimum number of points.

In contrast to many newer methods, it features a well-defined cluster model called “densityreachability”. Similar to linkage based clustering, it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects’ range. Another interesting property of DBSCAN is that its complexity is fairly low – it requires a linear number of range queries on the database – and that it will discover essentially the same results (it is deterministic for core and noise points, but not for border points) in each run, therefore there is no need to run it multiple times

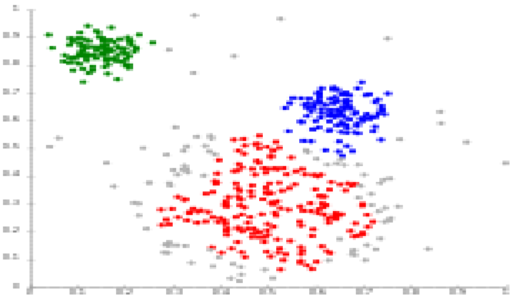

Figure 6

DENCLUE (DENSITY BASED CLUSTERING)

Denclueis a clustering method that depends upon density distribution function. DENCLUE uses a gradient hill-climbing technique for finding a local maxima of density functions . These local maxima are called density attractors, and only those local maxima whose kernel density approximation is greater than the noise threshold are considered in the cluster.

OPTICS

Ordering points to identify the clustering structure (OPTICS) is an algorithm for finding density-based

it addresses one of DBSCAN’s major weaknesses: the problem of detecting meaningful clusters in data of varying density. In order to do so, the points of the database are (linearly) ordered such that points which are spatially closest become neighbors in the ordering. Additionally, a special distance is stored for each point that represents the density that needs to be accepted for a cluster in order to have both points belong to the same cluster.

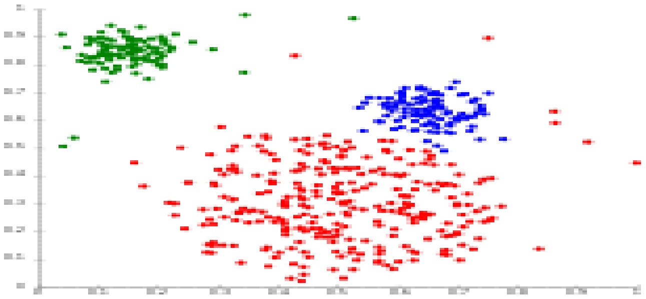

Figure 7

4.GRID BASED CLUSTERING

Grid-based clustering where the data space is quantized into finite number of cells which form the grid structure and perform clustering on the grids. Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids. Grid basedclustering is the fastest processing time that typically depends on the size of the grid instead of the data. The grid based methods use thesingle uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by thecell using a set of statistical attributes from the objects. These algorithms have a fast processing time, because they go through the data set once to compute the statistical values for the grids and the performance of clustering depends only on the size of the grids which is usually much less than the data objects. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE. All these methods use a uniform grid mesh to cover the whole problem. For the problems with highly irregular data distributions, the resolution of the grid mesh must be too fine to obtain a good clustering quality. A finer mesh can result in the mesh size close to or even exceed the size of the data objects, which can significant increase the computation load for clustering

- Creating the grid structure, i.e., partitioning the data space into a finite number of cells.
- Calculating the cell density for each cell.
- Sorting of the cells according to their densities.
- Identifying cluster centers.
- Traversal of neighbor cells.

ADVANTAGE

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

STING (Statistical Information Grid approach)

This approach breaks the available space of objects into cells of rectangular shapes in a hierarchy. It follows the top down approach and the hierarchy of the cells can contain multiple levels corresponding to multiple resolutions . The statistical information like the mean, maximum and minimum values of the attributes is pre computed and stored as statistical parameters, and is used for query processing and other data analysis tasks. The statistical parameters for higher level cells can be computed from the parameters for lower level cells. The complexity is O (K) where k denotes the total count of cells in last tier .

ADVANTAGES OF STING

- Query independent.
- The grid structure facilitates parallel processing and incremental updating.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600

http://www.ijcsjournal.com Volume 5, Issue 1, No 19, 2017 ISSN: 2348-6600
Reference ID: IJCS-247 PAGE NO: 1560-1571

Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

DISADVANTAGES OF STING

- The quality of the cluster can be degraded as the final cluster does not have any diagonal boundary.

COMPARISION OF THE VARIOUS CLUSTERING ALGORITHM

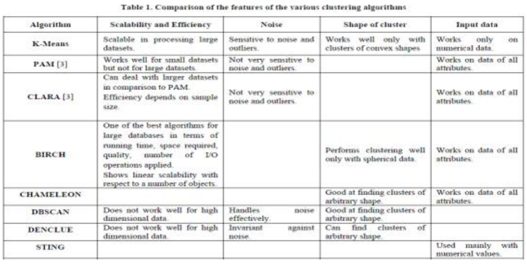

Table 1. Comparison of the features of the various clustering algorithms

Algorithm	Scalability and Efficiency	Noise	Shape of cluster	Input data
K-Means	Scalable in processing large datasets.	Sensitive to noise and outliers.	Works well only with clusters of convex shapes	Works only on numerical data.
PAM [3]	Works well for small datasets but not for large datasets.	Not very sensitive to noise and outliers.		Works on data of all attributes.
CLARA [3]	Can deal with larger datasets in comparison to PAM. Efficiency depends on sample size.	Not very sensitive to noise and outliers.		Works on data of all attributes.
BIRCH	One of the best algorithms for large databases in terms of running time, space required, quality, number of I/O operations applied. Shows linear scalability with respect to a number of objects.		Performs clustering well only with spherical data.	Works on data of all attributes.
CHAMELEON			Good at finding clusters of arbitrary shape.	Works on data of all attributes.
DBSCAN	Does not work well for high dimensional data.	Handles noise effectively.	Good at finding clusters of arbitrary shape.	
DENCLUE	Does not work well for high dimensional data.	Invariant against noise.	Can find clusters of arbitrary shape.	
STING				Used mainly with numerical values.

GRAPH-BASED CLUSTER

Graph Clustering is similar to a spectral clustering and it is Simple and scalable clustering method there are two types of graph clustering, they are Between-graph: clustering methods divide a set of graphs into different clusters and Within-graph: clustering methods divides the nodes of a graph into clusters. There are several algorithm for withingraph clustering. They are, Shared Nearest neighbour, Between-Centrality Based, Highly Connected Components, Maximum Clique Enumeration, Kernel K-means algorithm and finally Power Iteration cluster.

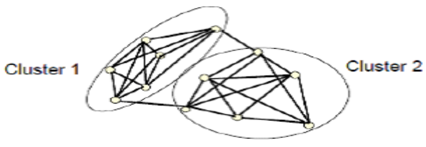

Figure 8

CONCLUSION

objective of datamining is extract the knowledge from large from large dataset. nd transform it into an understandable form for further use.clustering plays the vital role of datamining,data analyzing, Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Formation of clusters depend up on the algorithm used for clustering[1] , hierarchical clustering builds models based on distance connectivity, Partitioning is the centroid based clustering, the value of k-mean is set. Density based clusters are defined as area of higher density then the remaining of the data set. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells.clustering algorithms can be create either soft cluster(each object belongs to each cluster to a certain degree)or hard cluster (each object belongs to a cluster or not).clustering plays a promising role of natural grouping of objects,yet in future ample number of clustering algorithms forms the meticulous clusters.

REFERENCES

[1]AmandeepKaur Mann, NavneetKaur“ Survey Paper on Clustering Techniques” International Journal of Science, Engineering and Technology Research (IJSETR) April 2013

[2]PradeepRai, Shubha Singh” A Survey of Clustering Techniques” International Journal of Computer Applications,October 2010.

[3] OdedMaimon, LiorRokach, “DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK”, Springer Science+BusinessMedia.Inc, pp.321-352, 2005.

[4] David Pettinger and Giuseppe Di Fatta, “). “Scalability of Efficient Parallel K-Means”, IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 308-315.

[5]MadjidKhalilian, FarsadZamaniBoroujeni, Norwati Mustapha, Md. NasirSulaiman,. “K-Means Divide and

Conquer Clustering”, IEEE 2009, International Conference on Computer and Automation Engineering, pp. 306-309.

[6] Y.-T. Kao, E. Zahara, . A hybridized approach to data clustering, Expert Systems with Applications,

pp. 1754-1762.

[7] Nisha and Puneet Jai Kaur, “A Survey of Clustering Techniques and Algorithms”, IEEE (978-9-3805-4415-1), 2015

[8] Anoop Kumar Jain, Prof. Satyam Maheswari, “Survey of Recent Clustering Techniques in Data Mining”, International Journal of Computer Science and Management Research (2278-733X), Vol 1 Issue1.

[9] PavelBerkhin, “Survey of Clustering Data Mining Techniques”, Accrue Software, Inc.

[10] Namrata S. Gupta, Bijendra S. Agrawal, Rajkumar M. Chauhan, “Survey on Clustering Techniques of Data Mining”, American International Journal of Research in Science, Technology, Engineering & Mathematics (2328-3491), 9(3), December 2014-February 2015, pp. 206-211.