Data Mining using K-means and Hierarchical Clustering Algorithms: A Case Study Approach

Akshata Prabhu,
Electronics and Telecommunication,
K.J.Somaiya College of Engineering
Vidyavihar, Mumbai

Abstract— **Data mining is the processing of analyzing large datasets and database so has to derive patterns and establish relation between them. There are various disciplines available under Data mining and clustering is one of them. Clustering is the process of dividing data into groups that have similar objects. Each similar group is a cluster. The distance between the objects in the same group is minimum. It is possible to implement clustering algorithms via number of methods. This paper shows study and comparison between different clustering algorithms – Partition methods and hierarchical methods.**

Keywords— Website Analytics, K-means clustering, Hierarchical clustering

I. Introduction

Web analytics involves analyzing the behavior of individuals to a Web site. The process of analyzing and storing web utilization reports of consumers forms an important part of Web analytics. Whenever individuals visit website such as a mobile store or a cloth store, a web analytics software immediately stores measurable information about the user and the information related to users computer. The use of Web analytics enables business to attract new customers for goods or services and thereby increase the amount each customer spends. Web analytics is an inseparable part of customer relationship management analytics. Web Analytics include:-

a) Determining the chances a customer will repurchase a product

b) The amount of purchases made by each customer or group of customers

c) Observing the geographical regions from which the least and most purchases are made

d) Predicting which customers are likely to purchase in the future.

The main objective of Web Analytics is to promote products to consumers and improve the ratio of revenue to marketing costs. [1]

II. Data Mining

Data mining is a field that exploits statistical models, machine-learning and algorithms over the variants of data. Its main objective is to analyze different types of data (structured, unstructured and semi-structured) in order to gain insights from the data. Data mining is concerned with developing methods to explore the unique types of data and make the predictions for the future based on the trends in the past data.

The main steps in the process of Data Mining include:-

1. **Business understanding:-** This includes understanding the business objectives and finding out the business needs. It helps to consider the factors like resources, assumptions and constraints that are important in meeting the business needs.

2. **Data understanding:-** The data understanding phase includes collection of data from the authentic data sources. Once the data is collected it needs to be explored by tackling data mining questions which are addressed using querying, reporting and visualization.

3. **Data preparation:-** The data once collected needs to be selected, cleaned, constructed and formatted into the desired form.The exploration of data is conducted to derive hidden patterns from the data.

4. **Modeling:-** This includes selection of modeling techniques so that it can be applied on the prepared dataset. The dataset is divided into train and test sets. The models are build using the train dataset. The test scenarios must be generated to validate the quality and validity of the model. Models are evaluated to make sure that it meets the business requirements.

5. **Evaluation:-** In this phase, new business requirements may be raised due to the new patterns that have been observed in the data. Understanding the business value is an iterative process in data mining.

6. **Deployment:-**The stakeholders should be able to use the information and make decisions based on the data gained through the data mining process. The process of deployment can be simple or complex based the business requirements.

In this paper, we have studied the use of K-means clustering and Hierarchical-clustering approach to analyze the data related to web pages.

III. Algorithms Used

Clustering algorithms are unsupervised learning process. In clustering, the data objects having the same properties are grouped into a single cluster. The distance between the points in the same cluster should be minimum as compared to the data objects in different clusters. Clustering techniques are further divided into following categories: hierarchical clustering techniques and partitioning clustering techniques [2]. A short review of methods described below.

A. Partition Method

The simplest and easiest version of cluster analysis is partitioning, which organizes the object of set into group of clusters [2]. Simple example of partition methods is K-Means. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable k. The algorithm works iteratively to assign each data point to one of k groups based on the features provided. Data points are clustered based on feature similarity.

K-Means clustering can be implemented with the help of the steps described below:-

1) Initially choose k objects from the data and assign them as the cluster centers.

2) Consider the distance of each object from the cluster center. Assign the object to the cluster such that it has minimum distance. The distance between the points is calculated using Euclidean distance formula given below

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

3) Calculate the new position of each centroid by the mean value of object in cluster. [2]

4) The points stop moving at a certain stage and will get assigned to the cluster. Repeat steps 2 and 3 till the points stop moving.

For a given dataset, the minimum number of clusters cannot be specified. To overcome this, the results of multiple runs with different clusters is done and the best is chosen according to criteria.

There are many methods to determine the number of clusters but we have used elbow method. The idea behind clustering methods is to define clusters such that the total within-cluster sum of square (WSS) is minimized.

WSS is calculate by the formula :-

$$\sum_{j=1}^{k} \sum_{\forall x_i \in c_j} \|x_i - \mu_j\|^2, \qquad [3]$$

The total WSS measures the compactness of the clustering and it should be as small as possible. The Elbow method looks at the total WSS as a function of the number of clusters. One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

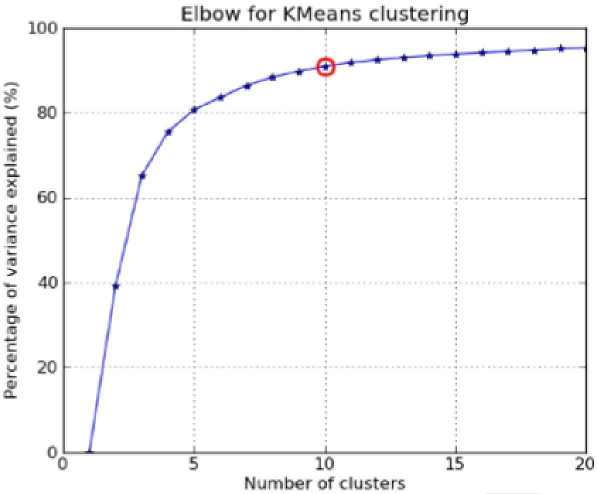


Figure I Plot of Number of Clusters and Percentage of Variance explained

From the above graph, we observe that at the bend, the WSS is reduced after a certain point. Hence we can chose the number of clusters as 10 in the above case.

B. Hierarchical Clustering

Hierarchical clustering is defined as the method in which clusters are formed in the form of a tree or hierarchy. Every node in the tree represents the different cluster and the clusters in the hierarchy are known as dendrograms [4]. Hierarchical clustering can be performed in two ways based on splitting and merging of clusters: divisive method and agglomerative method.

Agglomerative method works in the reverse direction of divisive method. In this method, the number of clusters are given initially and these clusters are merged in such a way that the two clusters to be merged are very similar to each other. These clusters are merged together until a large cluster is formed. Therefore, this method is also known as bottom-up approach. [4] [5]

Divisive method of hierarchical clustering is also known as top-down approach in which a large data set is given initially and this data set is further divided into a number of smaller subsets (known as clusters) until a threshold is reached. [4]

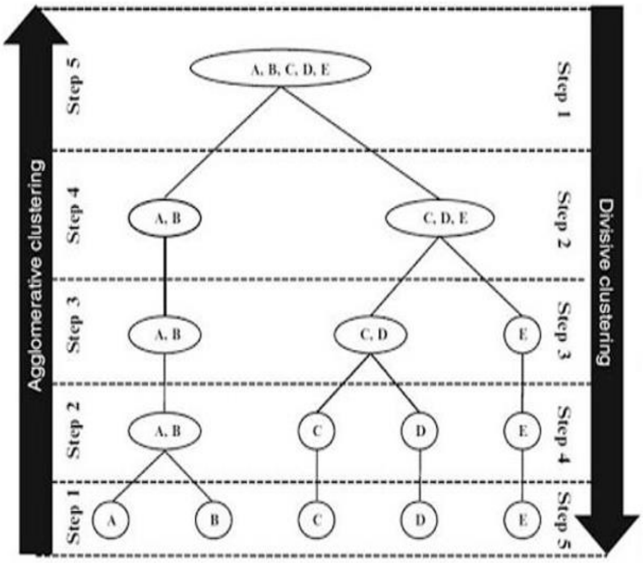


Figure II Dendogram for Agglomerative and Divisive Clustering

Types of Hierarchical Clustering :-

The Hierarchical Clustering is based on any one of the four methods.

Single-link Method: In Single Link Method, the distance between two clusters is the minimum of the distances between all pairs of patterns drawn one from each cluster.

Complete-link Method: In Complete-link Method,the distance between two clusters is the maximum of all pair wise distances between pairs of patterns drawn one from each cluster.

Average-link Method: In Average Link Method, the distance between two clusters is the average of all pair wise distances between pairs of patterns drawn one from each cluster

Centroid Method: In Centroid Method, the geometric center is computed. The distance between two clusters is equal to the distance between two centroids.

For the implemnatation of Hierarchical Clustering in our work, We have used Single Link Method.

IV. PROBLEM DEFINITION

Web Analytics is the key component of any E-commerce companies. It is helps the management team of an E-commerce company to understand the types of people who visit the website and make a purchase. Web analytics helps to believe that there are different market segments who make a higher purchase than the other segments and thereby create market for their customers. The following analysis has been made with the help of clustering techniques.

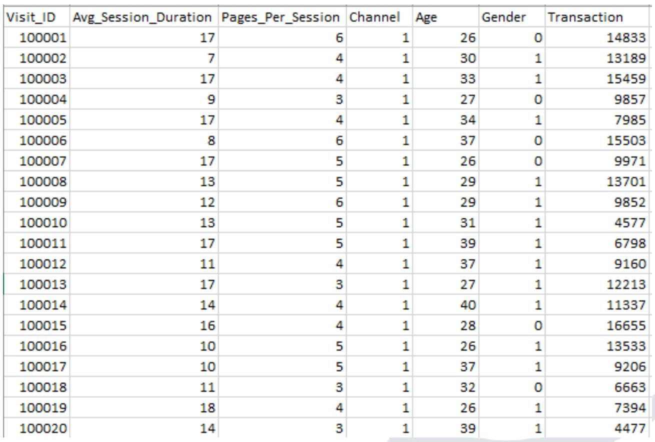
1. Similarity in the buying pattern of the people based on their demographics
2. Creating effective customer segments which help the management team in customer acquisitions

V. METHODOLOGY

This research applied two clustering techniques. They were K-means and Hierarchical Clustering. The Data was obtained from Google Analytics of the website, which includes variables about the customers. The analysis of the data was performed using R programming language using Rstudio Software. Rstudio is an open source software and is widely used for implementation of data mining algorithms and machine learning algorithms. The various attributes describing the consumers included:-

1) Visitor Id- Id number of the person visited

2) Average Session duration- Average number of minutes that particular customer has spent on the website.
3) Pages per session – Number of Pages visited per session
4) Channel – Paid channel 1, Organic channel – 0
5) Age- Age of the customer
6) Gender – Male-0, Female – 1
7) Transaction – Amount of purchase in rupees

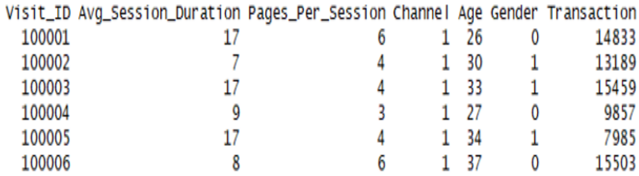


Visit_ID	Avg_Session_Duration	Pages_Per_Session	Channel	Age	Gender	Transaction
100001	17	6	1	26	0	14833
100002	7	4	1	30	1	13189
100003	17	4	1	33	1	15459
100004	9	3	1	27	0	9857
100005	17	4	1	34	1	7985
100006	8	6	1	37	0	15503
100007	17	5	1	26	0	9971
100008	13	5	1	29	1	13701
100009	12	6	1	29	1	9852
100010	13	5	1	31	1	4577
100011	17	5	1	39	1	6798
100012	11	4	1	37	1	9160
100013	17	3	1	27	1	12213
100014	14	4	1	40	1	11337
100015	16	4	1	28	0	16655
100016	10	5	1	26	1	13533
100017	10	5	1	37	1	9206
100018	11	3	1	32	0	6663
100019	18	4	1	26	1	7394
100020	14	3	1	39	1	4477

Figure III Dataset for Web Analytics

Steps in Data Preparation

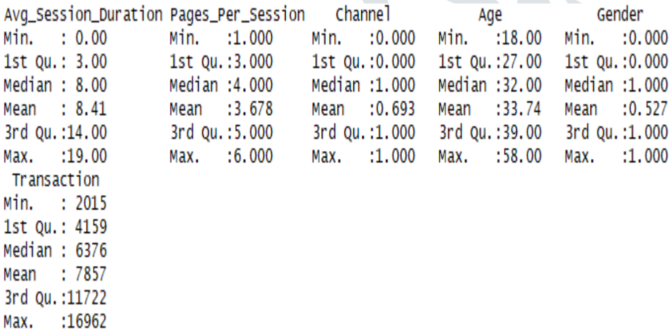
The first step in our approach is data selection and cleaning. In this step, we observed the structure of the data and checked if any missing values were present in the data. The unwanted variables which didn't affect the cluster formation were removed.



```
 Visit_ID Avg_Session_Duration Pages_Per_Session Channel Age Gender Transaction
   100001                   17                 6       1  26      0       14833
   100002                    7                 4       1  30      1       13189
   100003                   17                 4       1  33      1       15459
   100004                    9                 3       1  27      0        9857
   100005                   17                 4       1  34      1        7985
   100006                    8                 6       1  37      0       15503
```

Figure IV Structure of the data set

It was observed that the variable Visit_Id does not affect the analysis. Hence it was removed from the further analysis.



```
 Avg_Session_Duration Pages_Per_Session     Channel           Age            Gender
 Min.   : 0.00        Min.   :1.000     Min.   :0.000   Min.   :18.00   Min.   :0.000
 1st Qu.: 3.00        1st Qu.:3.000     1st Qu.:0.000   1st Qu.:27.00   1st Qu.:0.000
 Median : 8.00        Median :4.000     Median :1.000   Median :32.00   Median :1.000
 Mean   : 8.41        Mean   :3.678     Mean   :0.693   Mean   :33.74   Mean   :0.527
 3rd Qu.:14.00        3rd Qu.:5.000     3rd Qu.:1.000   3rd Qu.:39.00   3rd Qu.:1.000
 Max.   :19.00        Max.   :6.000     Max.   :1.000   Max.   :58.00   Max.   :1.000
  Transaction
 Min.   : 2015
 1st Qu.: 4159
 Median : 6376
 Mean   : 7857
 3rd Qu.:11722
 Max.   :16962
```

Figure V Summary of the data

After taking the summary of the data, it was observed that there are no missing values in the data.

VI. RESULTS AND ANALYSIS

A. *Hierarchical Clustering*

Figure III shows the hierarchical cluster formed on the website data.

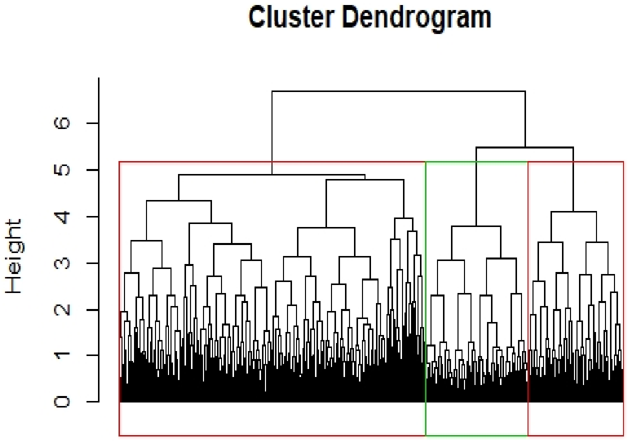


Figure III Dendrogram for Hierarchical clustering

The entire data was divided into three clusters. The data points were assigned to the cluster with minimum distance between them. The distance between the points was calculated using Euclidean distance formula. The method used for building the clusters was "complete method".

B. *K-Means Clustering*

The Website Data was divided into clusters using K-Means Clustering Algorithm. The first step in K-means clustering is to decide the number of clusters. Elbow Method was used to decide the number of Clusters. The Elbow Chart shows a plot Number of Clusters versus within group sum of square.

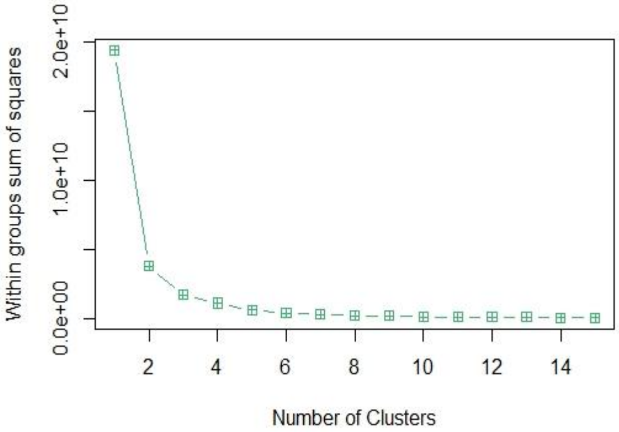
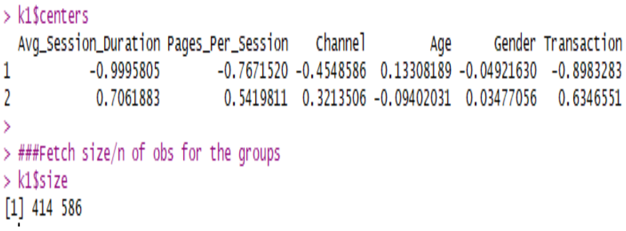


Figure IV Elbow Chart for deciding the number of Clusters

From the above graph we observed that there is significant change in the within groups sum of sqaure after k=2. Hence we have taken number of clusters as 2. We have also done the analysis using number of Clusters as 4.



```
> k1$centers
  Avg_Session_Duration Pages_Per_Session    Channel         Age      Gender Transaction
1           -0.9995805        -0.7671520 -0.4548586  0.13308189 -0.04921630  -0.8983283
2            0.7061883         0.5419811  0.3213506 -0.09402031  0.03477056   0.6346551
>
> ###Fetch size/n of obs for the groups
> k1$size
[1] 414 586
```

Figure V Centers and Size of the two Clusters

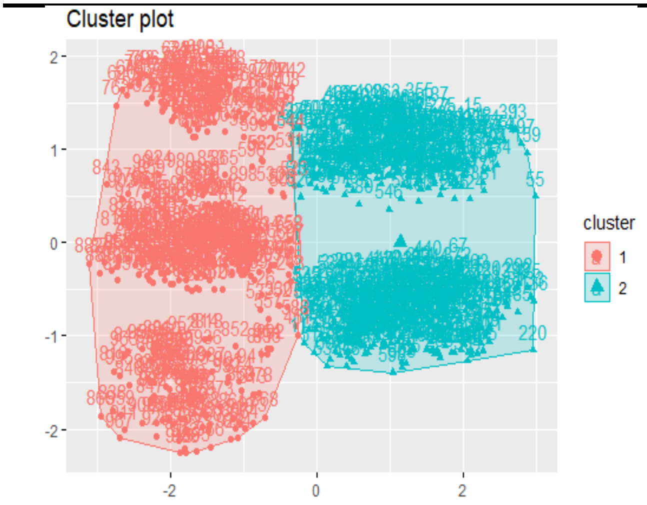
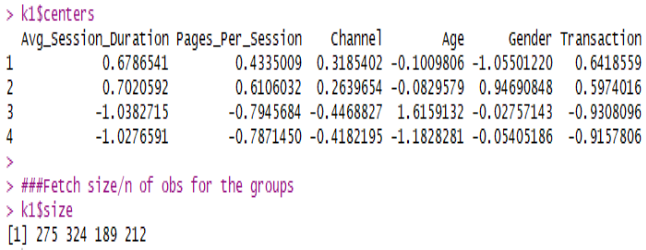


Figure VI Cluster Plot for Number of Clusters=2



```
> k1$centers
  Avg_Session_Duration Pages_Per_Session    Channel        Age      Gender Transaction
1            0.6786541         0.4335009  0.3185402 -0.1009806 -1.05501220   0.6418559
2            0.7020592         0.6106032  0.2639654 -0.0829579  0.94690848   0.5974016
3           -1.0382715        -0.7945684 -0.4468827  1.6159132 -0.02757143  -0.9308096
4           -1.0276591        -0.7871450 -0.4182195 -1.1828281 -0.05405186  -0.9157806
>
> ###Fetch size/n of obs for the groups
> k1$size
[1] 275 324 189 212
```

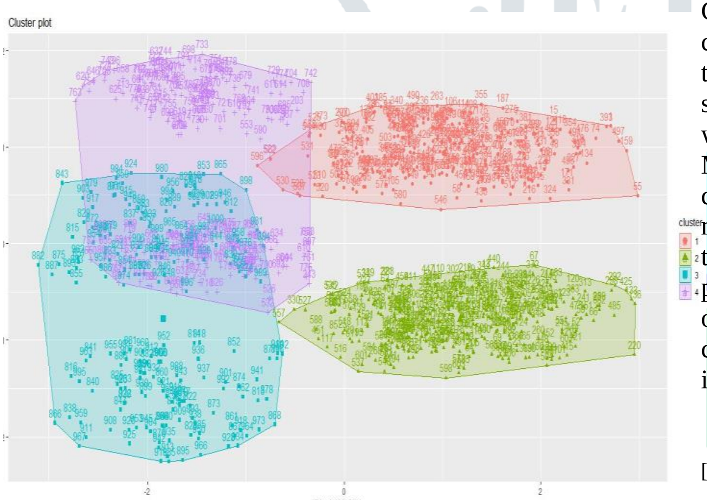


Figure VII Cluster Plot for Number of Clusters=4

C. Analysis of Clusters

From the figure below, We can clearly see a difference in the minimum, maximum and the average value of the variable transaction. A range of broader values are included in the cluster 1 where as smaller transaction values are included in cluster 2

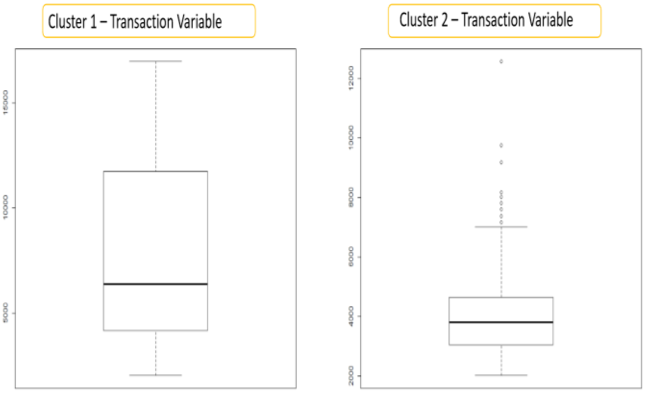


Figure VIII Boxplot for Transaction Variables for Number of Clusters = 2

It can be observed from the figure below that people in tha Age group from 25 to 40 have a higher purchasing value from the rest of the customers.

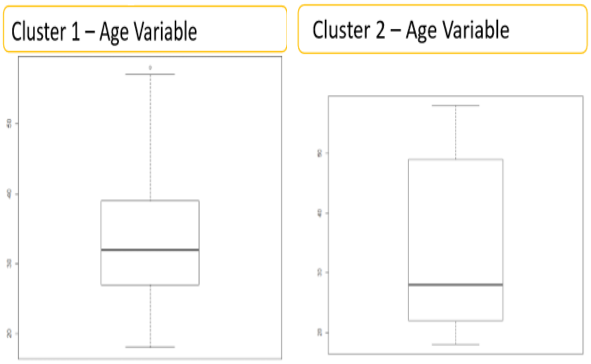


Figure VIII Boxplot for Age Variable for Number of Clusters = 2

VII. CONCLUSION

Based on the results obtained, Clustering Techniques were compared. It was observed that K-means Clustering can handle the big data well while Hierarchical clustering can't handle the big data well. K Means is found to work well when the shape of the clusters is hyper spherical while Hierarchical clustering is not found to work well when the shape of the clusters is hyper spherical. In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ while in Hierarchical Clustering the results are reproducible. K Means clustering requires prior knowledge of K i.e. no. of clusters we want to divide our data into But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram .

REFERENCES

[1] Anoopkumar M, Dr. A. M. J. Md. Zubair Rahman, "A Review on Data Mining Techniques and Factors Used in Educational Data Mining to Predict Student Amelioration,"

[2] Chintan Shah and Anjali Jivani " Comparison of Data Mining Clustering Algorithms", 2013 Nirma University International Conference on Engineering (NUiCONE)

[3] Nuwan Ganganath, Chi-Tsun Cheng, and Chi K. Tse, " Data Clustering with Cluster Size Constraints Using a Modified k-means Algorithm ", 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery

[4] Nisha and Puneet Jai Kaurt, " Cluster Quality Based Performance Evaluation of Hierarchical Clustering Method ", 2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015) Dehradun, India, 4-5 September 2015

[5] Zahra Nazari, Dongshik Kang & M.Reza Asharif, Yulwan Sung & Seiji Ogawa " A New Hierarchical Clustering Algorithm " , ICIIBMS 2015, Track2: Artificial Intelligence, Robotics, and Human-Computer Interaction, Okinawa, Japan .