Analysis of Complete-Link Clustering for Identifying and Visualizing Multi-attribute Transactional Data using MATLAB

Arna Prabha Jena
Dept. of Computer Science Engineering, Centurion University of Technology and Management, Bhubaneswar, India

Annan Naidu Paidi
Dept. of Computer Science Engineering, Centurion University of Technology and Management, Bhubaneswar, India

ABSTRACT
In recent years, entirely the data mining has drawn towards a great deal of interest in the field of information industry due to the wide availableness of enormous amount of data and the imminent need for turning such data into useful information and knowledge. Clustering is a powerful field of research in data mining. Many clustering algorithms have been developed to find patterns representing knowledge and are implicitly stored or captured in large databases etc, to provide decision support to the users. The quality of clustering can be assessed based on a metric of dissimilarity of objects, computed for various types of data. This paper presents, one of the agglomerative approaches of hierarchical clustering techniques i.e. complete-linkage clustering by considering four different types of distance metrics using Matlab toolbox, in order to compute distances (similarities/dissimilarities) between the new cluster and each of the old clusters.

Keywords
Data mining, cluster, clustering, hierarchical clustering, agglomerative, complete-linkage clustering, Matlab toolbox.

1. INTRODUCTION
An explosive development in transitory and stored data has generated a pressing need for new techniques and automated tools. These new techniques and tools further can intelligently assist us for transforming the large amount of data into useful information and knowledge. Data mining[1][2] can be viewed as a result of the natural growth of information technology. The database system industry has witnessed an evolutionary path in the development of the following functionalities: data collection and database creation, data management and advanced data analysis. Clustering [16] has been shown as one of the most commonly used data analysis techniques. So, clustering can be defined in terms of similarity i.e. the process of organizing data instances into groups whose members are similar in some manner. And eventually, a cluster is the collection of data instances which are common to each other and uncommon to the data instances in other clusters. Therefore, in the field of marketing the clustering is also called as segmentation (i.e partition into similar groups).

In recent years, data mining[5][7] is considered to be a multidisciplinary field, drawing work from areas including database technology, machine learning, statics, pattern recognition, information retrieval, artificial intelligence, high-performance computing, data visualization, etc.

This paper presents one of the data visualization approaches by using MATLAB. MATLAB[13] initiated as an interactive program for doing matrix calculations and has now extended to a high level mathematical language that can solve integrals and differential equations numerically and can plot a wide variety of graphs i.e it combines numeric computation, 2-D and 3-D graphics, and language capabilities in a single, easy-to-use environment. MATLAB is the foundation for all the MathWorks products. The MATLAB[14] application is built around the MATLAB language, and most use of MATLAB involves typing MATLAB code into the Command Window or executing text files containing MATLAB code and functions. So, MATLAB is both a language and a working environment, basically called as a matrix laboratory tool.

2. HIERARCHICAL CLUSTERING TECHNIQUES
Hierarchical clustering[4][9][10] techniques produce an embedded sequence of partitions. The result of hierarchical clustering[11][12] algorithm can be graphically displayed as tree, called Dendrogram[15], which graphically displays the merging process and the intermediate clusters or we can say how objects are grouped together step by step. For hierarchical clustering, this Dendrogram provides a taxonomy or hierarchical index or classification.

2.1 Working Principle/ Approach
There are two basic approaches for generating a hierarchical clustering:

2.1.1 Agglomerative
The agglomerative process is also known as bottom-up strategy/approach as one move up the hierarchy. Most of the hierarchical clustering methods belong to this category. It composed of the following steps-

1. Starts by placing each object in its own cluster.
2. At each step, it merges the most common or closest pair of clusters which requires a definition of cluster similarity or distance into larger and larger clusters.
3. Repeat, until all of the objects are in a single cluster.
4. Stop.

2.1.2 Divisive
It composed of the following steps-

1. Start with all objects in one cluster
2. At each step, it subdivides the cluster into smaller and smaller pieces.
3. Repeat, until only singleton (i.e. single) clusters of individual points remain.
4. Stop.

We can say that, it’s the reverse of agglomerative approach in which we need to decide, at each step, which cluster to split and how to split. This case follows top-down strategy/approach as one move down the hierarchy.

In this paper, we are considering the bottom-up strategy i.e agglomerative approach of hierarchical clustering techniques which is further classified based on linkage between clusters.

The agglomerative method erases rows and columns in the proximate matrix, as the old clusters are merged to form new ones. Suppose, the N*N proximate matrix is D = [d (x,y)]. The clusters are assigned with sequence numbers from 0, 1... (n-1) and L (k) is the level of the kth clustering. A cluster with sequence number s is denoted (s) and the proximate between clusters (p) and (q) is denoted as d [(p),(q)].

The following are the three mainly used types of agglomerative hierarchical clustering:

Single Linkage Clustering**:** Shortest/Minimum distance which uses the smallest distance between objects in the two clusters. This is the default one. The algorithm is composed of the following steps:

1. Start with the disjoint clustering having level L (0) =0 and sequence number s=0.
2. Find the least dissimilar pair of clusters in the current clustering i.e the pair (p) and (q) in such a manner that d [(p),(q)] = min d [(x),(y)], where the minimum is considered from all pairs of clusters in the current clustering.
3. Increment the sequence number: s = s +1. Merge clusters (p) and (q) into a single cluster to form the next clustering s. Set the level of this clustering to L(s) = d [(p),(q)].
4. Update the proximate matrix, D, by deleting the rows and columns corresponding to clusters (p) and (q) and adding a row and column corresponding to the newly formed cluster. The proximate between the new cluster, denoted as (p,q) and old cluster (k) is defined as: d [(k), (p,q)] = min [d[(k), (p)], d [(k),(q)]].
5. If all objects are in one cluster, stop. Else, go to step 2.

Average Linkage Clustering**:** Unweighted average distance, also known as group average which uses the average/mean distance between all pairs of objects in any two clusters. The algorithm is composed of the following steps:

1. Start with the disjoint clustering having level L (0) =0 and sequence number s=0.
2. Find the least dissimilar pair of clusters in the current clustering i.e the pair (p) and (q) in such a manner that d [(p),(q)] = min d [(x),(y)], where the minimum is considered from all pairs of clusters in the current clustering.
3. Increment the sequence number: s = s +1. Merge clusters (p) and (q) into a single cluster to form the next clustering s. Set the level of this clustering to L(s) = d [(p),(q)].
4. Update the proximate matrix, D, by deleting the rows and columns corresponding to clusters (p) and (q) and adding a row and column corresponding to the newly formed cluster. The proximate between the new cluster, denoted as (p,q) and old cluster (k) is defined as: d [(k), (p,q)] = mean [d[(k), (p)], d [(k),(q)]].
5. If all objects are in one cluster, stop. Else, go to step 2.

Complete Linkage Clustering**:** Furthest/Maximum distance which uses the largest distance between objects in the two clusters. This paper explains the complete linkage[8] type which is also known as diameter method. In complete linkage type of clustering, the distance between one pair of nearest cluster is highest to another pair of nearest cluster i.e the distance between two clusters is determined by the most distant nodes in the two clusters. At each step the complete linkage clustering algorithm tends to minimize the increase in diameter of the clusters. If the true cluster are compact or closely packed and approximately of equal size then the method will produce high-quality clusters. The algorithm is composed of the following steps:

1. Start with the disjoint clustering having level L (0) =0 and sequence number s=0.
2. Find the least dissimilar pair of clusters in the current clustering i.e the pair (p) and (q) in such a manner that d [(p),(q)] = min d [(x),(y)], where the minimum is considered from all pairs of clusters in the current clustering.
3. Increment the sequence number: s = s +1. Merge clusters (p) and (q) into a single cluster to form the next clustering s. Set the level of this clustering to L(s) = d [(p),(q)].
4. Update the proximate matrix, D, by deleting the rows and columns corresponding to clusters (p) and (q) and adding a row and column corresponding to the newly formed cluster. The proximate between the new cluster, denoted as (p,q) and old cluster (k) is defined as: d [(k), (p,q)] = max[d[(k), (p)], d [(k),(q)]].
5. If all objects are in one cluster, stop. Else, go to step 2.

To be simple, here is the flowchart for the complete view of the Agglomerative Hierarchical clustering process which further includes four types of distance calculation measures to determine the different distance matrix:.

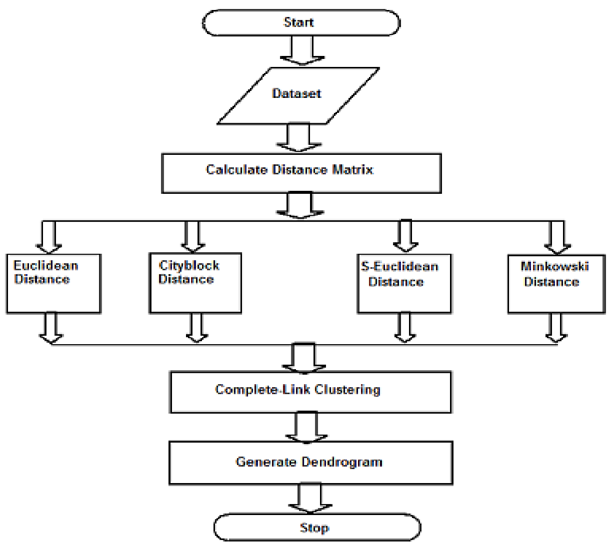

Fig 1: Agglomerative Hierarchical Clustering Process

Distance between each node can be calculated by the following distance measures:

Euclidean Distance Formulae:

$$d(x, y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad (1)$$

Cityblock (Manhattan) Distance Formulae:

$$d(x, y) = \sum_{i=1}^{m} |x_i - y_i| \quad (2)$$

Standardized Euclidean Distance Formulae:

$$d(x, y) = \sqrt{\sum_{i=1}^{m}\left(\frac{x_i}{s_{x_i}} - \frac{y_i}{s_{y_i}}\right)^2} \quad (3)$$

Where, s_{x_i} and s_{y_i} are the standard deviation.

Minkowski Distance Formulae:

$$d(x, y) = \{\sum_{i=1}^{m} |x_i - y_i|^g\}^{1/g} \quad (4)$$

When, g = 1, the Minkowski distance gives the City Block distance.

g = 2, the Minkowski distance gives the Euclidean distance (default).

g = ∞, the Minkowski distance gives the Chebychev distance.

3. HIERARCHICAL CLUSTERING IN MATLAB

In Matlab, the Statistics Toolbox incorporates the “pdist” for calculating pairwise distance between observations, “linkage” which creates a hierarchical cluster tree using the specified algorithm and “dendrogram” generates a dendrogram plot of the hierarchical binary cluster tree by the linkage function. Here, a complete description is explained followed by an example.

Syntax:

Y = pdist(X)
Y= pdist(X,metric)
Z = linkage(Y)
Z = linkage(Y,method)
Z = linkage(Y,method,metric)
[...] = dendrogram(...,'colorthreshold',t)

Description

Y= pdist(X) computes the Euclidean distance between pairs of objects in n-by-p data matrix X. Rows of X correspond to observations; columns correspond to variables. Y is a row vector of length n(n–1)/2, corresponding to pairs of observations in X.

Y = pdist(X,metric) computes the distance between objects in the data matrix, X, using the method specified by metric, which can be any of the following character strings (e.g. 'Euclidean', 'Seuclidean', 'Cityblock', 'Minkowski', etc.)

Z = linkage(Y) creates a hierarchical cluster tree from the distances in Y.

Z = linkage(Y,method) creates the tree using the specified method and these methods differ from one another in how they measure the distance between clusters(e.g. 'single', 'complete', 'average', etc.).

Z = linkage(Y,method,metric) performs clustering using the distance measure metric to compute distances between the rows of Y.

[...] = dendrogram(...,'colorthreshold',t) assigns a unique color to each group of nodes in the dendrogram where the linkage is less than the threshold t. t is a value in the interval [0,max(Z(:,3))]. Setting t to the string 'default' is the same as t = .7(max(Z(:,3))). 0 is the same as not specifying 'colorthreshold'. The value max(Z(:,3)) treats the entire tree as one group and colors it all one color.

An example 1: In the following example nine clusters are computed from the U.S Cities data [3] which is the distance in miles of nine areas of U.S city using Complete- linkage, and see how linkage is calculated by applying one of the distance formulae i.e Euclidean distance in Matlab. Here, we are considering numeric series (1-9) instead of the nine U.S city areas name as shown in the table1 & 2:

Table 1. U.S City Data Set

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
B0S	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Table 2. U.S City Data Set (Contd…)

	1	2	3	4	5	6	7	8	9
1	0	206	429	1504	963	2976	3095	2979	1949
2	206	0	233	1308	802	2815	2934	2786	1771
3	429	233	0	1075	671	2684	2799	2631	1616
4	1504	1308	1075	0	1329	3273	3053	2687	2037
5	963	802	671	1329	0	2013	2142	2054	996
6	2976	2815	2684	3273	2013	0	808	1131	1307
7	3095	2934	2799	3053	2142	808	0	379	1235
8	2979	2786	2631	2687	2054	1131	379	0	1059
9	1949	1771	1616	2037	996	1307	1235	1059	0

By applying Euclidean distance formulae, the following input distance matrix is obtained:

Table 3. Distance Matrix Using Euclidean Distance Metric

	1	2	3	4	5	6	7	8	9
1	0	555	1021	2544	2429	6549	6926	6589	4655
2	555	0	560	2529	2112	6506	6863	6504	4465
3	1021	560	0	2376	1781	6364	6700	6321	4212
4	2544	2529	2376	0	2854	6076	6285	5869	4351
5	2429	2112	1781	2854	0	5174	5485	5100	2757
6	6549	6506	6364	6076	5174	0	1408	1776	3074
7	6926	6863	6700	6285	5485	1408	0	792	3189
8	6589	6504	6321	5869	5100	1776	792	0	2769

```
load citydata
b=round(pdist(a,'euclidean'));
cs=squareform(b);
c=linkage(b,'complete');
firstfive = round(c(1:5,:)); % first 5 rows of c
figure,d=dendrogram(c,'colorThreshold','default');
firstfive=
1    2    555
7    8    792
3    10   1021
6    11   1776
5    12   2429
```

Where, firstfive argument gives the first five calculated values by applying Complete- Linkage and 'a' computes the distance between objects in the data matrix (i.e. U.S City Data).

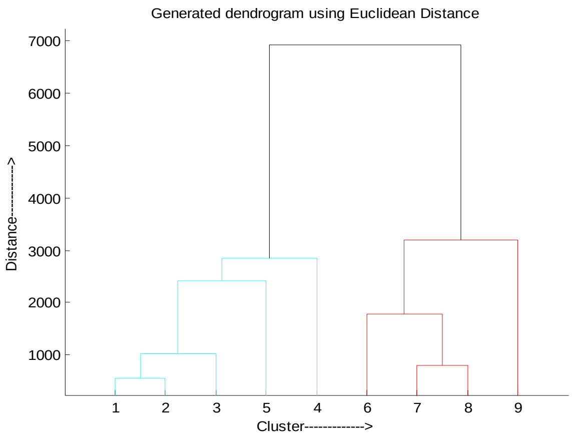

Fig 2: Hierarchical Clustering (Complete-link) for U.S City Data Using Euclidean Distance

Initially, by taking the input distance from Table 2 the Euclidean distance for all pair of cities has been calculated as shown in Table 3. The nearest pair of cities is BOS and NY i.e 1 and 2 according to Table 3, at distance 555. These are merged into a single cluster called "1/2" in the first step. Then we compute the distance from this new compound object to all other objects. In complete- link clustering the rule is that the distance from the compound object to another object is equal to the longest distance from any member of the cluster to the outside object. So the distance from "1/2" to 3 is chosen to be 1021, which is the distance from 2 to 3. Similarly, the distance from "1/2" to 9 is chosen to be 4655. Likewise, we have to continue this process until only one cluster is left. Finally, the last clusters are merged at level 6926. Lastly, according to the merged cluster and their associated values the dendrogram tree will be obtained. Simillarly, distance-matrix and dendrogram trees are generated for Complete-Linkage clustering method by considering the four different types of distance metrics as shown:

```
b1=round(pdist(a,'euclidean'));
c=linkage(b1,'complete');
figure,d=dendrogram(c,'colorThreshold','default');
```

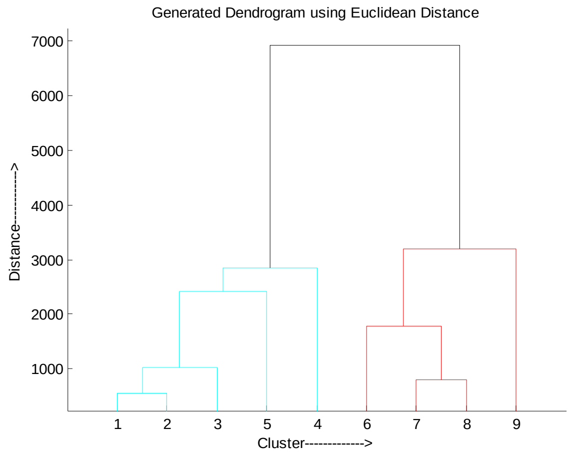

Fig 3: Hierarchical Clustering (Complete-link) for U.S City Data Using Euclidean Distance

```
b2=round(pdist(a,'seuclidean'));
c=linkage(b2,'complete');
figure,d=dendrogram(c,'colorThreshold','default');
```

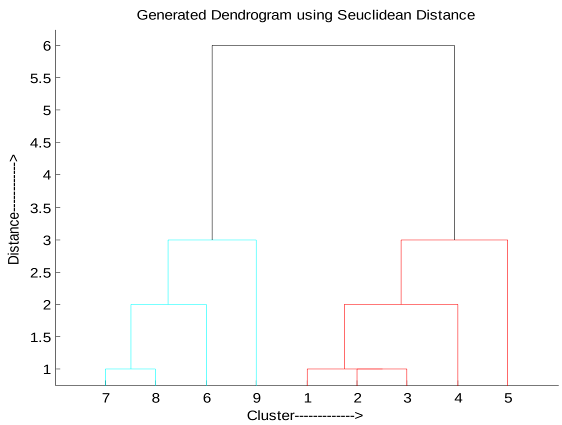

Fig 4: Hierarchical Clustering (Complete-link) for U.S City Data Using Seuclidean Distance

```
b3=round(pdist(a,'cityblock'));
c=linkage(b3,'complete');
figure,d=dendrogram(c,'colorThreshold','default');
```

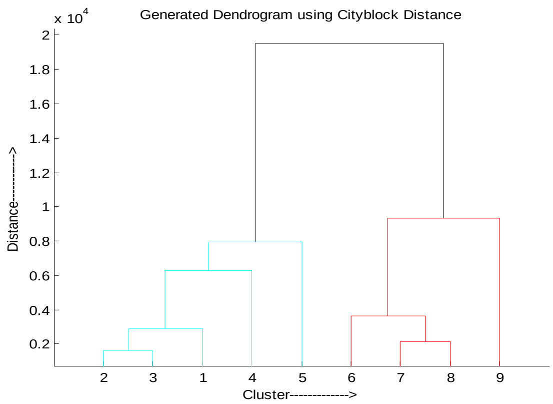

Fig 5: Hierarchical Clustering (Complete-link) for U.S City Data Using Cityblock Distance

```
b41=round(pdist(a,'minkowski',1));
c=linkage(b41,'complete');
figure,d=dendrogram(c,'colorThreshold','default');
```

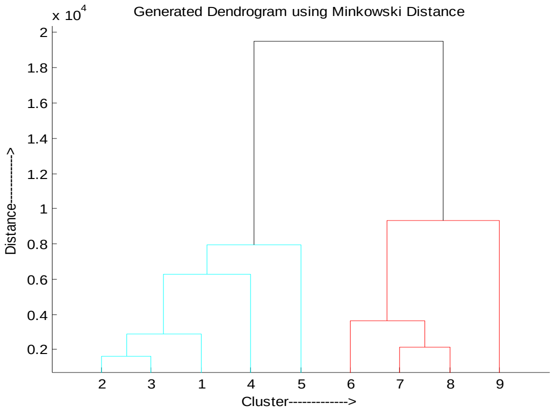

Fig 6: Hierarchical Clustering (Complete-link) for U.S City Data Using Minkowski Distance (g=1)

```
b42=round(pdist(a,'minkowski',2));
c=linkage(b42,'complete');
figure,d=dendrogram(c,'colorThreshold','default');
```

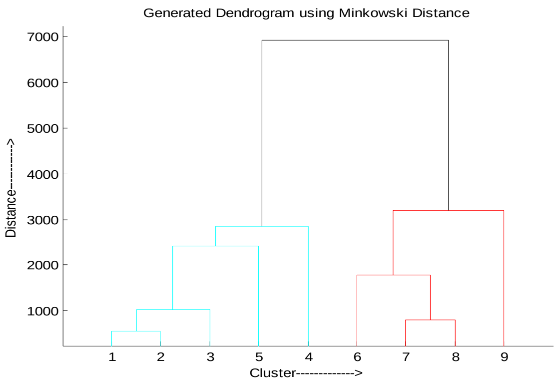

Fig 7: Hierarchical Clustering (Complete-link) for U.S City Data Using Minkowski Distance (g=2)

```
b43=round(pdist(a,'minkowski',3));
c=linkage(b43,'complete');
figure,d=dendrogram(c,'colorThreshold','default');
```

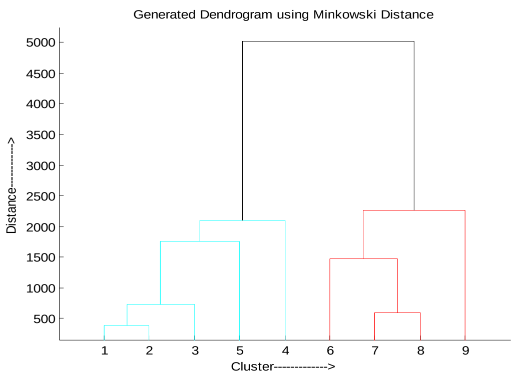

Fig 8: Hierarchical Clustering (Complete-link) for U.S City Data Using Minkowski Distance (g=3)

An example 2: In the following example, out of nine five clusters are computed from the Hospital dataset consisting of six parameters (i.e. Name, Sex, Age, Weight, Smoker, Blood pressure). Complete-linkage clustering is achieved on two parameters (i.e. Weight and Blood pressure) and ignoring other all parameter information. Here, Linkage and Dendrogram tree is generated by using one of the distance formulae i.e Euclidean distance formula in Matlab:

Table 4. Hospital Data

Name	Sex	Age	Weight	Smoker	Blood Pressure	
Smith	Male	38	176	TRUE	124	93
Johnson	Male	43	163	FALSE	109	77
Williams	Female	38	131	FALSE	125	83
Jones	Female	40	133	FALSE	117	75
Brown	Female	49	119	FALSE	122	80
Davis	Female	46	142	FALSE	121	70
Miller	Female	33	142	TRUE	130	88
Wilson	Female	40	180	FALSE	115	82
Moore	Male	28	183	FALSE	115	78

```
load hospital
hospital(1:5,1:6)
X = [hospital.Weight hospital.BloodPressure];
firstfive1 = round(X(1:5,:)); % first 5 rows of X
b=round(pdist(firstfive1,'euclidean'));
cs=squareform(b);
c=linkage(b,'complete');
firsttwo = round(c(1:2,:)); % first 2 rows of c
figure,d=dendrogram(c,'colorThreshold','default');
```

```
firstfive1=
176    124    93
163    109    77
131    125    83
133    117    75
119    122    80
```

Where, firstfive1 argument gives the first five weight and blood pressure values respectively from Hospital dataset.

```
firsttwo=
3    4    11
5    6    16
```

Where, firsttwo argument gives the first two calculated values by applying from Hospital dataset by applying Complete-Linkage and returns the dendrogram as an output. By applying Euclidean distance formulae, the following Dendrogram is obtained:

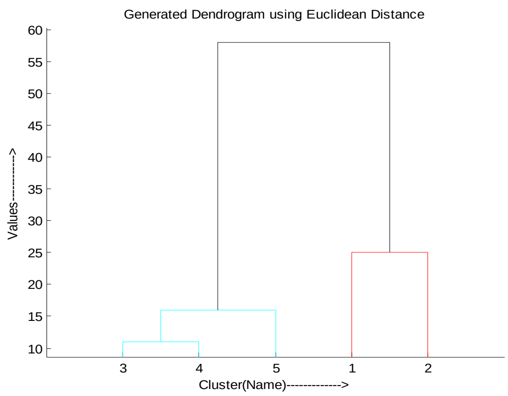

Fig 9: Hierarchical Clustering (Complete-link) for hospital Data

Likewise, more Linkage and Dendrogram tree can generated by including all the distance metrics using Matlab. The main advantage is, at each step the complete linkage clustering algorithm tends to minimize the increase in diameter of the clusters. If the true cluster are compact or closely packed and approximately of equal size then the method will produce high-quality clusters. Furthermore, by running all the distance metrics in one flow, we can test and find out the Cityblock Distance as one of the best and less time consuming distance metrics.

4. FUTURE SCOPE

This work provides many directions for future research. There are many other clustering algorithms that are worthwhile and useful in various applications. So, we can apply the above mentioned four distance metrics on other clustering techniques to view the result, to check the performance. Furthermore, image processing clustering analysis can also be done by taking image data points using different clustering techniques.

5. CONCLUSION

This paper presents, one of the agglomerative approaches of hierarchical clustering technique i.e complete linkage clustering by taking U.S City data and Hospital data as an example. In this technique, the distance between one pair of nearest cluster is maximum to another pair of nearest cluster i.e the distance between two clusters is determined by the most distant nodes in the two clusters and due to which, at each step the complete linkage clustering algorithm tends to minimize the increase in diameter of the clusters. If the true cluster are compact or closely packed and approximately of similar size then the method will produce high-quality clusters.

Furthermore, in order to calculate linkage we have included four more types of distance metrics for calculating and generating corresponding distance matrix and the associated dendrogram tree respectively. So, the main benefit is, by giving one dataset as an input we are obtaining different dendrogram tree as an output and because of that we can have different point of views with different trees. Even though it is not possible to describe all the algorithms and step by step calculation in detail, but we have tried to discuss the hierarchical clustering methods, their properties and algorithms using MATLAB. The example discussed and provided references will give the interested reader a rough idea of MATLAB tools that are helpful in further research.

6. ACKNOWLEDGMENT

The successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragement crown all the efforts with success. I would like to convey my heart full thanks to my guide Asst. Prof. AnnanNaidu Paidi, for his guidance and support in every step of this paper. I would like to convey my sincere thanks to all the faculty and friends who directly or indirectly helped me for the successful completion of this paper.

7. REFERENCES

[1] Adomavicius Gediminas, Bockstedt Jesse "*C-TREND: Temporal Cluster Graphs for Identifying and Visualizing Trends in Multiattribute Transactional Data*", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,, vol. 20, no.6,pp.721-733, June 2008.

[2] Ahmed Riaz Syed, "*Application of data mining in retail business*", Proc. International Conference on Information Technology: Coding and Computing, April 2004.

[3] Rani Radha D., Bharati Vini A., Sravani A., "*Analysis of Dendrogram Tree for Identifying and Visualizing trends in Multi attribute Transactional Data*", International Journal of Engineering Trends and Technology, vol. 3, pp. 14-18, [2012].

[4] Sridath Phani V.R, Srinivas Kudipudi and Rao Srinivasa V., "*Visualization of Time-Series Cluster Graphs Using Hierarchical Clustering Techniques*", International Journal of Engineering Sciences and Technologies, vol.7, pp.65-69, [2011].

[5] Ruey-shun Chen, Ruey-chyi Wu and J. Y. Chen, "*Data Mining Application In Customer Relationship Management Of Credit Card Business*", Proceedings Of The 29th Annual International Computer Software And Applications Conference, pp.1-2, Taiwan, [2005].

[6] Daniel A. Keim, And Hans-peter Kriegel ,"*Visualization Techniques For Mining Large Databases: A Comparison*", IEEE Transactions On Knowledge And Data Engineering, Vol. 8, No. 6, pp.923-938, December 1996.

[7] Jia-dong Ren, Jie Bao, Hui-w Huang, "*The Research On Spatio-temporal Data Model And Related Data Mining*", proceedings Of The Second International Conference On Machine Learning And Cybernetics,, pp.37.-40,November 2003.

[8] Jiawei Han, Micheline Kamber, "*Data Mining Concepts and Techniques*", Elsevier Inc, 2006.

[9] Ying Peng, Yongyi Ma, Huairong Shen,"*Clustering Belief Functions using Agglomerative Algorithm*",Information Engineering and Computer Science(ICIECS), pp. 1-4,[2010].

[10] Takumi, satoshi,miyamoto,sadaaki,"*Top-Down Vs Bottom-Up methods of linkage for asymmetric Agglomerative Hierarchical clustering*",Granular computing(GrC), pp. 459-464,[2012].

[11] Srinivas M, Mohan C.K, "*Efficient clustering Approach using incremental and hierarchical clustering methods*",neural Networks(IJCNN),International Joint conference Publication, pp.1-7,[2010].

[12] Hakim R.B.F., Subanar, winarko E., "*Reducing Dendrogram Instability of Features using Rough Set indiscernibility Level*",Distributed Framework and Applications, pp. 1-10,[2010].

[13] Sigmon Kermit, "*MATLAB Primer*", Third Edition, 1993.

[14] Sony.Krotha Rachitha, Merugula Suneetha, "*A Brief Survey on Document Clustering Techniques Using MATLAB*", International Journal of Computer & Organization Trends (IJCOT), pp. 1-6,[2013].

[15] Jena, A.P., Naidu A., "*Analysis of Complete-Link Clustering for Identifying and Visualizing Multiattribute Transactional Data* ", International Journal of Computer Science & Engineering Technology (IJCSET), Vol.4, pp.850 – 854, 07 Jul 2013.

[16] Liu Bing, "*Web Data Mining*", Springer International Edition.