An empirical study on shortest path for Graph clustering in Network analysis

*Charan Kumar G.
Research scholar, Department of Mathematics, Srikrishnadevaraya University, Anantapuramu, (A.P.), India-515003.

G. Shobha Latha
Professor, Department of Mathematics, Srikrishnadevaraya University, Anantapuramu, (A.P.), India-515003.

Corresponding author: charankumarganteeda@gmail.com

Abstract. Network theory deals with the study of graphs in which nodes connected by branches. It helps to determine the shortest route between two places, time schedule for the activities of a project and minimum cost flow in pipeline networks. In the present paper, we with the existing algorithms based on network theory and finding the shortest path from use the cluster algorithms to find the can be compare one to another and also shortest paths when the clusters are grouping based on cluster analysis the results which we obtain observed that the similarity levels of single, average and complete linkage methods. By applying the existing algorithms based on graph theory can analyze the shortest path from one to another where as clustering provides the shortest path as well as the similarity index for standardized and non-standardized variables.

Keywords: Minimal spanning tree, Shortest route, Dijkstra’s algorithm, clustering, Dendogram.

1. Introduction
Graph theory widely used in computer science for studying the algorithms such as Dijkstra’s, Kruskal’s and Prim’s. Computer programs have been used to represent large graphs occurred in PERT, flow problems, transportation networks, electrical networks, circuit layouts. A graph can be used to represent almost any physical situation involving discrete objects and a relationship among them. Routes between the cities can be represented by using graphs. Characterization of the graphs is based on their structures. An algorithm is a step by step procedure will lead to the solutions of the problem. Ore, O (1961) identified a problem regarding the tracing of graphs. Charnes and Cooper (1967)
studied some network characterization for mathematical programming and accounting applications to planning and control. As point out by Knuth (1968), page 4, every algorithm must have five important features: finiteness, definiteness, input, output and effectiveness. Evans and Minieka (1992). Optimized algorithms for networks and graphs. Robinson et al (1993) studied an integrated design distribution system at dowbrands. Birch (1997) introduced a New Data Clustering Algorithm and Its Applications. Russell R. Barton (2015) developed simulation Metamodelling. Saurav Kaushik (2016) studied clustering and its different methods. Erica Flapan et al (2017) developed spatial graphs to intrinsic knotting and linking results. Kamaldeep singh et al (2017) studied spatial modeling of urban road traffic using graph theory. A matrix is a convenient and useful way of representing a graph. In many applications of graph theory, such as in electrical network analysis and operations research, matrices also turn out to be the natural way of expressing the problem. Praveen and Rama (2017) studied A K- Means Clustering algorithm on Numeric data.

In the present study, following the work of Dijkstra’s algorithm and Praveen and Ram (2017) we observed the shortest routes by applying shortest route algorithms and clustering algorithms which have been provided in the above literature have been used for finding the shortest path from one to another. Our study mainly focused on network theory shortest paths when the clusters are grouping based on cluster analysis the results which we obtain observed that the similarity levels of single, average and complete linkage methods. Here we used the existing algorithms based on graph theory can analyze the shortest path from one to another where as clustering provides the shortest path as well as the similarity index for standardized and non-standardized variables.

### 2. Graph theoretic approach in networks

A distance between two vertices u and v of a connected graph is the length of the shortest path connecting them. For a connected graph G.

- $E(v) = \max \text{ dist}(v,x)$ the eccentricity of v in G.
- $D(G) = \max E(v)$ is the diameter of a G
- $R(G) = \min E(v)$ the radius of G.
2.1 Definition of Network: A Network consists of a set of nodes linked by arcs. The notation for describing a network is \((N, A)\), where \(N\) is the set of nodes, and \(A\) is the set of arcs. A network is said to be connected if every two distinct nodes are linked by at least one path. A tree is a cycle free connected network comprised of a subset of all the nodes, and a spanning tree is a tree that links all the nodes of the network. The minimal spanning tree links the nodes of a network using smallest total length of connecting branches. If we consider, the application of roads linking towns, either directly or passing through other towns. The minimal spanning tree solution provides the most economical design of the road system. The shortest-route problem determines the shortest route between a source and destination in a transportation network. Dijkstra’s algorithm provides the shortest routes between the source node and every other node in the network.

2.2 Dijkstra’s algorithm: (Ref. 6, Page No. 247-248): Let \(u_i\) be the shortest distance from source node 1 to node \(i\), and define \(d_{ij}(\geq 0)\) as the length of the arc \((i, j)\). The algorithm defines the label for an immediately succeeding node \(j\) as

\[
[u_j, i] = [u_i + d_{ij}, i], d_{ij} \geq 0
\]

The label for the starting node is \([0, -]\), indicating that the node has no predecessor. Node labels in Dijkstra’s algorithm are of two types: temporary and permanent. A temporary label at a node is modified if a shorter route to the node can be found. Otherwise, the temporary status is changed to permanent.

1. Label the source node (node 1) with the permanent label \([0, -]\). Set \(i = 1\).
2. General step i:
   i) Compute temporary labels \([u_i, d_{ij}, i]\) for each node \(j\) with \(d_{ij} > 0\), provided is not permanently labeled. If node \(j\) already has an existing temporary label \([u_j, k]\) through another node \(k\) and if \(u_i + d_{ij} < u_j\), replace \([u_j, k]\) with \([u_i + d_{ij}, i]\).
   ii) If all the nodes have permanent labels, stop. otherwise select the label \([u_r, s]\) having the shortest distance (\(=u_r\)) among all temporary labels (break ties arbitrarily). Set \(i = r\) and repeat step i.
3. Clustering Approach in Networks:

The method of identifying similar objects of data in a dataset is called clustering. The quality of a clustering is based on the similarity measure used by the method and its implementation. Similarity is expressed in terms of a distance function and is denoted by \( d(i,j) \).

**Single Linkage:** It is a smallest distance between points and also known as nearest neighborhood method.

\[
D_{SL}(C_i, C_j) = \min d(a,b); \ a \in C_i, \ b \in C_j
\]

**Complete Linkage:** largest distance between points and also known as maximum method.

\[
D_{CL}(C_i, C_j) = \max d(a,b); \ a \in C_i, \ b \in C_j
\]

**Average linkage:** Average distance between points.

\[
D_{AL}(C_i, C_j) = \frac{1}{n_i n_j} \sum d(a,b); \ a \in C_i, \ b \in C_j
\]

**Ward’s method:** It begins with one cluster for each individual sample. At each iteration, among all pairs of clusters, it merges the pair that produces the smallest squared error for the resulting set of clusters.

**Dendrogram:** a tree data structure which illustrates hierarchical clustering techniques. Each level shows clusters for that level.

**Example:** To illustrate the various hierarchical clustering algorithms we shall use the sample data similar to Praveen and Rama (2010) A K-Means Clustering algorithm on numeric data that consists of 5 two-dimensional points, which are given in table: 1. The x and y coordinates of the points related to the road network data which gives the distances in miles between pairs.
of cities are mentioned in table 2. Here we use Dijkstra’s algorithm to find that shortest route between first city and every other city in the network.

| x | 4 | 8 | 15 | 24 | 24 |
|---|---|---|----|----|----|
| y | 4 | 4 | 8  | 4  | 12 |

**Table: 1**

The distance matrix for the data given in the clustering analysis is

|   | 1  | 2  | 3  | 4  | 5  |
|---|----|----|----|----|----|
| 1 | -  | 4  | 11.7| 20 | 21.5|
| 2 | 4  | -  | 8.1 | 16 | 17.9|
| 3 | 11.7| 8.1| -  | 9.8| 9.8 |
| 4 | 20 | 16 | 9.8 | -  | 8  |
| 5 | 21.5| 17.9| 9.8| 8  | -  |

**Table: 2**

By applying Dijkstra’s algorithm to the above data we obtain the following graph with desired route is A→B→C→D with a total of length of 29.9 miles and also observed that the maximum degree of graph is ‘4’, Graph radius is 11.7 and graph diameter is 21.5.
Cluster analysis: Consider the data set with ‘5’ clusters each consisting of one sample.

\[
\begin{array}{c}
x: \quad 4 & 8 & 15 & 24 & 24 \\
y: \quad 4 & 4 & 8 & 4 & 12 \\
\end{array}
\]

Distance between cluster Centroids

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
 & Cluster1 & Cluster2 & Cluster3 & Cluster4 & Cluster5 \\
\hline
Cluster1 & 0.0000 & 0.4391 & 1.6455 & 2.1953 & 3.1336 \\
Cluster2 & 0.4391 & 0.0000 & 1.3566 & 1.7562 & 2.8433 \\
Cluster3 & 1.6455 & 1.3566 & 0.0000 & 1.4919 & 1.4919 \\
Cluster4 & 2.1953 & 1.7562 & 1.4919 & 0.0000 & 2.2361 \\
Cluster5 & 3.1336 & 2.8433 & 1.4919 & 2.2361 & 0.0000 \\
\hline
\end{array}
\]

After applying clustering analysis to the above data we get the distance between cluster Centroids. For the same set of data we apply Dijkstra’s algorithm we obtain the following graph with desired route is A→B→C→D with a total of length of 4.78 miles and also observed that the maximum degree of graph is ‘4’, Graph radius is 1.65 and graph diameter is 3.13.
4. Hierarchical Cluster Analysis of Observations:

1. Euclidean distance and single linkage

| Step | Number of clusters | Similarity level | Distance level | Clusters joined | Number of observations in new cluster |
|------|--------------------|------------------|----------------|----------------|---------------------------------------|
| 1    | 4                  | 85.99            | 0.439          | 1              | 2                                     |
| 2    | 3                  | 56.71            | 1.357          | 1              | 3                                     |
| 3    | 2                  | 52.39            | 1.492          | 1              | 5                                     |
| 4    | 1                  | 52.39            | 1.492          | 1              | 4                                     |

2. Euclidean Distance, Complete Linkage

| Step | Number of clusters | Similarity level | Distance level | Clusters joined | Number of observations in new cluster |
|------|--------------------|------------------|----------------|----------------|---------------------------------------|
| 1    | 4                  | 85.99            | 0.439          | 1              | 2                                     |
| 2    | 3                  | 52.39            | 1.492          | 3              | 5                                     |
| 3    | 2                  | 29.94            | 2.195          | 1              | 4                                     |
| 4    | 1                  | 0                | 3.134          | 1              | 3                                     |

3. Euclidean Distance, Average Linkage

| Step | Number of clusters | Similarity level | Distance level | Clusters joined | Number of observations in new cluster |
|------|--------------------|------------------|----------------|----------------|---------------------------------------|
| 1    | 4                  | 85.99            | 0.439          | 1              | 2                                     |
| 2    | 3                  | 52.39            | 1.492          | 3              | 5                                     |
| 3    | 2                  | 40.51            | 1.864          | 1              | 4                                     |
| 4    | 1                  | 31.23            | 2.155          | 1              | 3                                     |

4. Euclidean Distance, Ward Method

| Step | Number of clusters | Similarity level | Distance level | Clusters joined | Number of observations in new cluster |
|------|--------------------|------------------|----------------|----------------|---------------------------------------|
| 1    | 4                  | 85.99            | 0.439          | 1              | 2                                     |
| 2    | 3                  | 52.39            | 1.492          | 3              | 5                                     |
| 3    | 2                  | 36.56            | 1.988          | 3              | 4                                     |
| 4    | 1                  | -12.23           | 3.517          | 1              | 3                                     |
Dendogram- Single linkage method:

![Dendogram- Single linkage method](image)

Dendogram- Complete Linkage Method

![Dendogram- Complete Linkage Method](image)
Dendogram-Average linkage method:

![Dendogram-Average linkage method diagram]

Dendogram-Ward Method:

![Dendogram-Ward Method diagram]
5. Comparison of Hierarchical clustering algorithms:

| Step | Standardized Similarity level | Non-standardized Similarity level |
|------|-------------------------------|-----------------------------------|
|      | Single linkage | Complete linkage | Average linkage | Ward linkage | Single linkage | Complete linkage | Average linkage | Ward linkage |
| 1    | 85.99           | 85.99              | 85.99              | 81.43       | 81.43          | 81.43          | 81.43          |
| 2    | 56.71           | 52.39              | 52.39              | 62.86       | 62.86          | 62.86          | 62.86          |
| 3    | 52.39           | 29.94              | 40.51              | 54.28       | 54.28          | 54.28          | 51.42          |
| 4    | 52.39           | 0                  | 31.23              | -12.23      | 54.28          | 0              | 26.34          | -31.34       |

6. Result and conclusions:

By applying Dijkstra’s algorithm to the data with desired route is A→B→C→D with a total of length of 29.9 miles and also observed that the maximum degree of graph is ‘4’, Graph radius is 11.7 and graph diameter is 21.5. Before converting the distance matrix we apply the clustering algorithm and determine the distance matrix between cluster Centroids and we got desired route is A→B→C→D with a total of length of 4.78 miles and also observed that the maximum degree of graph is ‘4’, Graph radius is 1.65 and graph diameter is 3.13. Hence, we observed that both the methods follow the same pattern even though clustering will help us when we have large data. When we are dealing large data clustering will help us in finding the shortest routes. We also observed that the similarity levels of three methods are at 1st all have same similarity, 2nd, 3rd and 4th showed different similarity levels for standardized and Non-standardized variables which are represented in table and graphs. When we consider the large data we are unable to identify the spatial variations by using graph theory in such cases we use the technique of clustering and finding the shortest routes by using different approaches. When we are dealing with the spatial data there may be the existence of autocorrelation, it is difficult to identify the similar groups in such cases clustering will help to find the distance matrix of similar and dissimilar entities. It will be very much helpful in finding the shortest routes as well as similarity index and the spatial relationships between the entities.
7. References
1. www.pearsonhighered.com/taha
2. Birch (1997). A New Data Clustering Algorithm and Its Applications. Data Mining and Knowledge Discovery, Volume 1, Issue 2.
3. Charnes, A., and W. Cooper, “Some Network Characterization for Mathematical Programming and Accounting applications to planning and control”, The Accounting Review, Vol. 42, No. 3, PP. 24-52.
4. Erica Flapan, Thomas W. Mattman, Blake Mellor, Ramin Naimi, and Ryo Nikkuni, RECENT DEVELOPMENTS IN SPATIAL GRAPH THEORY, Contemporary Mathematics, vol. 689, 2017, pp. 81-102.
5. Evans, J., and E. Minieka, Optimization algorithms for networks and graphs, 2nd Edition, Marcel Dekker, Newyork, 1992.
6. Hamdy, A. Taha (2014). Operation Research: An Introduction, Pearson, 9th edition, Thomson Press.
7. Kamaldeep Singh Oberoi, Geraldine Del Mondo, Yohan Dupuis and Pascal Vasseur (2017), Spatial modeling of urban road traffic using graph theory, SAGEO-Rouen, 6-9.
8. Knuth, D. E. (1968). The art of computer programming, Fundamental algorithms, Addison-Wesley publishing company, Inc., Reading, Mass Vol. 1.
9. Narsingh Deo (1984), Graph theory with applications to engineering and computer science, Prentice hall of India private limited, New Delhi.
10. Ore, O (1961). “A Problem regarding the tracing of graph”, Rev. elementary Math., Vol. 6, 49-53.
11. Robinson, E., L. Gao, and S. Muggenborg (1993), “Designing an Integrated distribution system at Dowbrands, Inc”, Interfaces, Vol.23, No.3, PP. 107-117.
12. Russell R. Barton (2015). Tutorial: simulation metamodelling, Proceedings of the 2015 Winter Simulation Conference L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.
13. Saurav Kaushik (2016). An Introduction to clustering and different methods of clustering.
14. https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf
15. P. Praveen and B. Rama (2017). A K- Means Clustering algorithm on Numeric dat, International journal of Pure and applied mathematics, Vol. 117, No. 7, P. No. 157-164.

Examples taken from published papers: -
  i) P. Praveen and B. Rama (2017). A K- Means Clustering algorithm on Numeric data, International journal of Pure and applied mathematics, Vol. 117, No. 7, P. No. 157-164.
  ii) https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf