aurum

AURUM MÜHENDİSLİK SİSTEMLERİ VE MİMARLIK DERGİSİ
AURUM JOURNAL OF ENGINEERING SYSTEMS AND ARCHITECTURE

Cilt 2, Sayı 1 | Bahar 2018
Volume 2, No. 1 | Spring 2018, 49-66

IMPLEMENTATION OF DIFFERENT CLUSTERING ALGORITHMS

Amna Mohamed M. ABURAS[1]
[1]Department of Electrical and Electronics Engineering
Faculty of Engineering and Natural Sciences, Altınbaş University-Turkey
amna.aburas@ogr.altinbas.edu.tr

Wamidh MAZHER[2]
[2]Department of Electrical and Electronics Engineering
Faculty of Engineering and Natural Sciences, Altınbaş University-Turkey
wamidh.mazher@ogr.altinbas.edu.tr

Osman Nuri UÇAN[3]
[3]Department of Electrical and Electronics Engineering
Faculty of Engineering and Natural Sciences, Altınbaş University-Turkey
osman.ucan@altinbas.edu.tr

Oğuz BAYAT[4]
[4]Department of Electrical and Electronics Engineering
Faculty of Engineering and Natural Sciences, Altınbaş University-Turkey
oguz.bayat@altinbas.edu.tr

Abstract

Spectral clustering is developed for both normalized and unnormalized methods. However, selecting between the two methods is not established in the GUI (Graphical User Interface) yet . In this paper , we implement different clustering algorithms using GUI-MATLAB, then, the clustering by these three methods, is compared for similar pairs of datasets. Our model is employing such three different clustering methods which are spectral, hierarchical and density based methods, then employing different geometrical, multi-range, and multi-level similar datasets pairs of graph for clustering. As result, the above three clustering algorithms are experimented for different environments which are (geometrical, multi-range and multi-level). The simulation result shows the clustering of these pairs of geometrical datasets which are: Concentric circles, Semi-circles, and Aggregation. Accordingly, the spectral algorithm has superior clustering in case of big datasets more than 2000 pairs points and range more than 500 levels among datasets.

Keywords*: Clustering, K-means, Spectral method, Laplace, Eigenvector, GUI.*

FARKLI SINIFLANDIRMA ALGORİTMALARININ UYGULAMALARI

Özet

Spektral kümeleme hem normalize hem de normalize edilmemiş yöntemler için geliştirilmiştir. Bununla birlikte, iki yöntem arasında seçim yapmak henüz GUI'de (Grafik Kullanıcı Arayüzü) kurulmamıştır. Bu yazıda, GUI-MATLAB kullanarak farklı kümeleme algoritmaları uyguluyoruz, daha sonra bu üç yöntemle kümeleme, benzer veri

kümeleri çiftleri için karşılaştırılıyor. Modelimiz, spektral, hiyerarşik ve yoğunluk temelli yöntemler gibi üç farklı kümeleme yöntemini kullanmaktadır, daha sonra kümeleme için farklı geometrik, çok aralıklı ve çok düzeyli benzer veri kümeleri grafikler kullanmaktadır. Sonuç olarak, yukarıdaki üç kümeleme algoritması, (geometrik, çok menzilli ve çok seviyeli) farklı ortamlar için denenmiştir. Benzetim sonucu, bu çift geometrik veri kümelerinin kümelenmesini göstermektedir: Eş merkezli daireler, yarı daireler ve toplama. Buna göre, spektral algoritma, veri kümeleri arasında 2000'den fazla çift nokta ve 500'den fazla veri kümesindeki üstün kümeleme özelliklerine sahiptir.

Anahtar Kelimeler: Sınıflandırma, K-ortalama, Spektral metod, Laplace, Özdeğer vektör, GUI.

1. INTRODUCTION

Clustering analysis groups are known to be; one of the multivariate statistical analysis methods, whose units are not known precisely, that helps to distinguish subgroups (groups, classes) whose variables are similar. The basic purpose of the clustering analysis is to group the units based on their characteristic features. Clustering analysis is one of the analysis methods that have been on the agenda in recent years [1],[2]. This method is interpretable and effective that is why it is used in many science and business fields like, genetics, statistics, biology, medicine, agriculture, veterinary, engineering, economic and administrative sciences, technical sciences, social sciences, data mining etc. . In the scientific studies, clustering is employed for classification purposes.

In clustering analysis, both the clusters number and structure of heterogeneous data are investigated. In clustering, the aim is to group the similar objects in the same group (cluster) and put the different objects in other groups (clusters). Clustering analysis is essentially different from parsing analysis. While observations in the decomposition (parsing) analysis are predefined and broken up into known numbers, the clustering analysis does not contain any preliminary information about the number or the structure of the groups. In some methods, clustering begins by finding similarities between all pairs of observations, where sometimes similarities are found based on distance among objects. In other clustering methods, the choice of cluster centers or a comparison between intra-cluster (within one cluster) and inter-cluster (among clusters) is made.

Applying clustering analysis categorized in four steps. In the first step, observations that are not classified in the population are converted into data matrix which consists units and variables (Unit $\times$ Variable). Determining the similarities/differences of the units and variables are done in the second step, such as Euclidean, Manhattan, Pearson, etc.

Similarities /differences among matrix elements which determine the distance and closeness among units, is obtained by using one of the similarity/distance measures. In the third step, Units/variables in the similarity/difference matrix with one of the methods are grouped. The clusters formed by the appropriate measures and methods are interpreted in the fourth step. The most critical aspect of the clustering analysis is how determining the number of clusters. The researcher needs to minimize the attribute before deciding the number of clusters. The correct selection of the cluster number governs the quality of the clusters to be created. Several methods have been developed to determine the appropriate number of clusters.

Although different methods are used to determine the number of clusters, but the researcher's knowledge level and the professional experience are important to obtain meaningful results. Based on the existence of variant clustering methods that employed different distance measures, the researcher can be confused to choose the appropriate method. For this reason, clustering analysis is one of the most used methods in recent years because it is based on utilizing the basic components. In this approach, it is possible to reduce the number of variables (in case of a large number of variables) and draw detailed information from the observation results (score) for the first two basic components.

2. CLUSTERING METHODS

For clustering, there are many methods to do that [3], [4]. In this paper, our study is focus on three clustering methods and compression the clustering among them by those three selected methods which are, spectral, hierarchical, and density methods.

2.1 Hierarchical Methods

Hierarchical methods are based on grouping objects into tree structures called Dendrograms. The methods of building construction are examined in two parts: (i) Coupler clustering and (ii) Parsing clustering. Hierarchical methods do not need k value, but they need to determine a threshold to specify when to stop the tree structure creation. The algorithms using hierarchical methods are listed as: (i) Combining and parsing algorithms. Which are divided into: (a) aggregation clustering AGNES and (b) decomposition clustering DIANA, (ii) BIRCH, (iii) CURE, and (iv) CHAMELEON[4].

The Hierarchical clustering methods perform hierarchical decomposition of the units in the data set by using the distance values of the units to each other. During hierarchical decomposition, a tree diagram known as a dendogram is employed. The tree diagram provides clusters visualization obtained by hierarchical clustering. The number of clusters is visually decided.

In the grouping hierarchical method, each unit or each observation is initially considered as a set. Then the two closest clusters or observations are assembled in a new cluster. This reduces the number of clusters in each step. The divisor hierarchical method starts with a large cluster of all observations. By eliminating similar observations, smaller clusters are created. Each observation is continued until a single cluster is formed [4].

The hierarchical clustering method can be expressed by an algorithm consisting of four steps: (i) the process starts with n individuals n sets, (ii) the nearest two sets of digits are combined to the smallest digits, (iii) the cluster number is a reduced recursive distance matrix, and (iv) Steps 2 and 3 are repeated n-times [5].

Trying different clustering methods in the analysis can help verifying the results. Depending on the properties of the peers, some clustering methods may create more appropriate clusters than others. The seven most commonly used hierarchical clustering methods are as follows: (i) Single Linkage Method (Single Linkage), (ii) Complete Linkage Method, (iii) Average Linkage Method, (iv) McQuitty Link Clustering

Method (v) Global (Central) Link Clustering Method (Centroid Linkage), (vi) Median Linkage Method, and (vii) Ward Link Clustering Method (Ward Linkage) [6].

2.2 Density Based Clustering Methods

In large data sets, you need to divide the data into smaller sub-sets. If the database is large, a good classification algorithm is required. Classification is done by clustering method which collects similar data into different clusters. On the other hand, using clustering can cause some problems: it is often difficult for users to know which input parameters should be used for a given database. Finally, the shapes of the clusters can be ambiguous and can be complicated in bad situations.

2.2.1 The DBSCAN Algorithm

The DBSCAN algorithm can identify clusters in large spatial data sets. For this reason, little information is needed about the domain. The process is fast and well scaled.

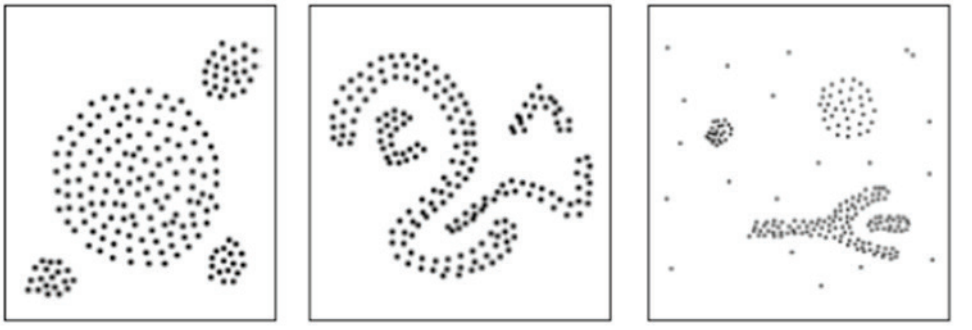

Figure 1. The node distribution of their different databases, taken from SEQUOIA 2000 benchmark database[7].

The following section will describe further how the DBSCAN algorithm works. Its computing process is based on six rules or definitions, creating two lemmas.

Definition 1:(Directly density-reachable)There are two kinds of points belonging to a cluster; they are border points and core points, as can be seen in Figure 2-a.

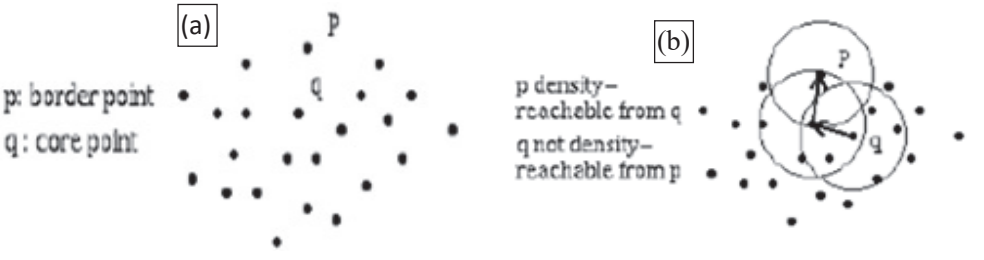

Figure 2. (a) Border- and core points, (b) Point p is density-reachable from point q and not vice versa [2].

Definition 2: (Density-reachable)"A point p is density-reachable from a point q with respect to Eps and Minp^{st} if there is a chain of points p1...,pn, p1=q, pn=p suchthat pi+1 is directly density-reachable from pi." , as shown in Figure 2-b.

2.2.2 Comparisons (DBSCAN vs. CLARANS)

The DBSCAN algorithm is compared with another clustering algorithm. This is called CLARANS. This is an enhancement of k-medoid algorithms [8]. In comparison with K-medoid, CLARANS works with databases for about a thousand objects [9],[10]. The classification of the CLARANS and DBSCAN algorithms are shown in Figures 3, where their run times are specified in Figure 4.

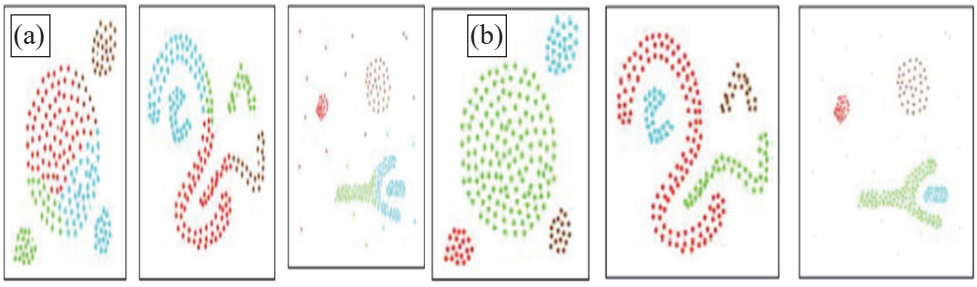

Figure 3. The classification of the CLARANS and DBSCAN algorithms [2].

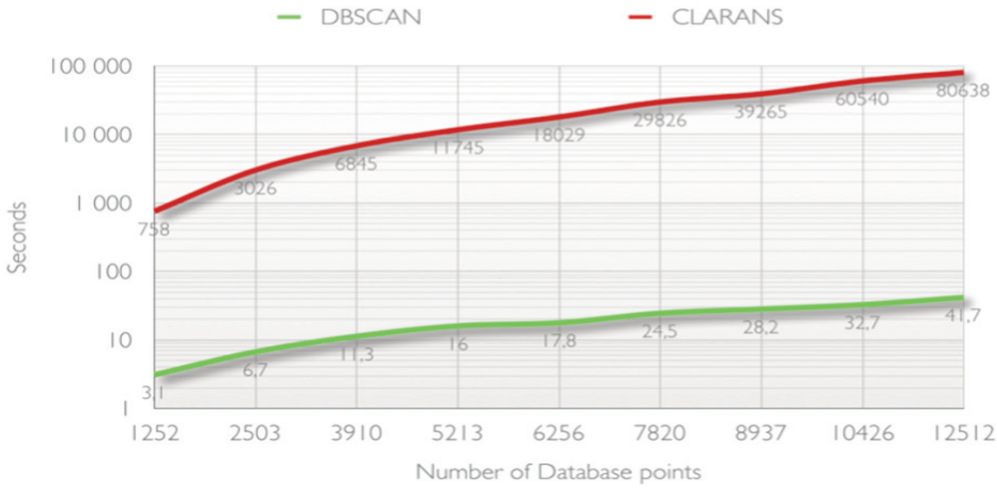

Figure 4. The Run time in seconds [2].

2.3 Spectral clustering

Spectral clustering is known to be; a clustering technique that is used for: "image processing", "data mining", and "machine learning communities". The deterministic polynomial can be considered to be superior

to the traditional clustering algorithms such as the K-median from some angles, to the ability to model clustering without clustering. For example, it has the ability to capture clusters as shown in Figure5-b of the spectral clustering, whereas the K-tool will fail Figure5-a.

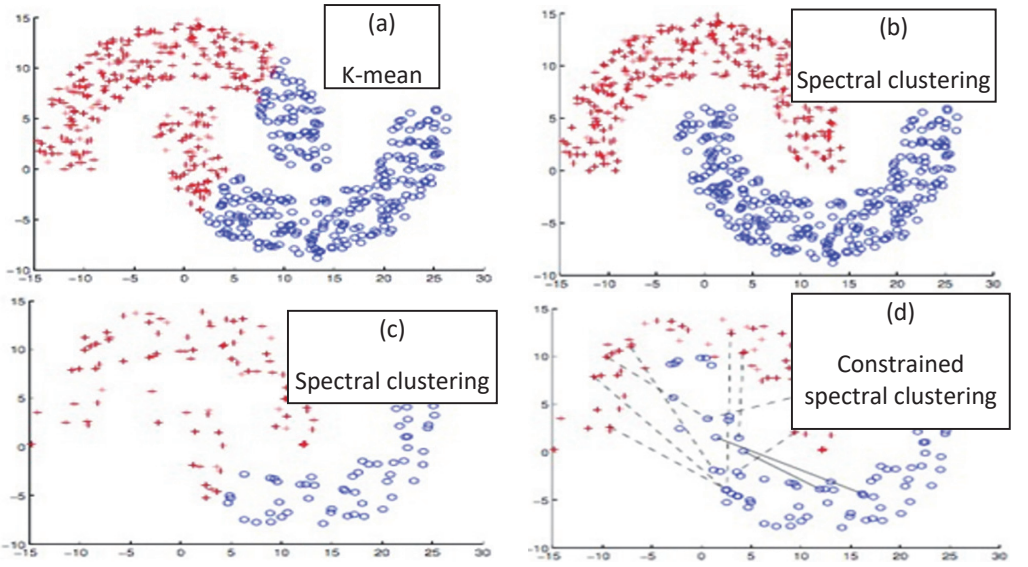

Figure 5. Example of variant clustering [2] types.

The advantage of spectral clustering is approved in most applications. Spectral clustering was initially suggested to point out to an uncontrolled learning problem where the data samples are untagged and all available information is coded in the graphic Laplace. Also, when using the same toy dataset as shown in Figure 5-c, the sets become so sparse that the separation becomes isolated from them. We can formulate it into various forms to be free of the unwanted part of the spectral cluster. Figure 5 shows examples of different clustering types: Based on how the constraints are applied, the restricted spectral clustering can be divided into two categories. The first category handles the graph directly according to Laplace, wherein the second category; the constraints are utilized to specify the possible solution space.

Terminologies and Notations:

A collection with N instances of data is modeled by an undirected, weighted graph G (V, E, A) , where each instance of data corresponds to a vertex (node) inV, E is the edge set and A is the associated with affinity matrix, which is "symmetric" and "non-negative"[11]. The diagonal matrix (D_{ii}) is known as the matrix degree of G graph, the diagonal (D11... DNN) is given by Eq (1) as follows:

$$D_{ii} = \sum_{j=1}^{N} A_{ij} \quad (1)$$

then

$$L = D - A\text{-}\mathbf{1}. \quad (2)$$

Table 1 of Notations

Symbol	Meaning
G	undirected (weighted) graph
A	The affinity matrix
D	The degree of matrix
I	identity matrix
L/$\bar{L}$	"unnormalized" and normalized Laplacian graph
Q/$\bar{Q}$	unnormalized/normalized constraint of matrix
Vol	volume of Ggraph

L is called the unnormalized graph Laplacian of G [12]. Assuming G is connected (i.e. any node is reachable from any other node), L has the following properties:

3. Similarity Graphs Construction

The following steps will be implemented in the application regardless of the input dataset:

First we shall compute the similarity matrix S**:** this is done by calculating Euclidean distances between point pairs and then applying the Gaussian kernel.

Second step is constructing the similarity graph: This step can be done by applying one of the following techniques:

1- Complete (Fully Connected) Similarity Graph:

The most trivial technique which keeps the similarity matrix S as it is. All points with positive similarity are connected with each other and all edges are weighed by sj. The width of the neighborhoods is controlled by t a parameter.

2- e-neighborhood Similarity Graph:

In which, we define an e and nominate all points with pairwise distances smaller than e to be connected. As the distances are more or less than the same this method will yield an unweight graph[13] ,[14].

3- k-nearest neighborhood Similarity Graph:

In this method, we define a vertex, and connect the k number of nearest neighboring vertices. As it is self-implying, this method will yield a non-symmetric neighborhood and lead to a directed graph. After we get rid of this directionality, we can weigh the edges by the similarity of adjacent points [15]. Third step is derivation of Laplace matrices with subsequent projection to the space of their k smallest eigenvectors: We derive the Laplacian Lsyrn and then extract the first k eigenvectors of this matrix. Then we form the matrix that contains the eigenvectors and normalize the rows.

The last step is applying k-means algorithm on the output of the previous step: the intended output will be k number of clusters.

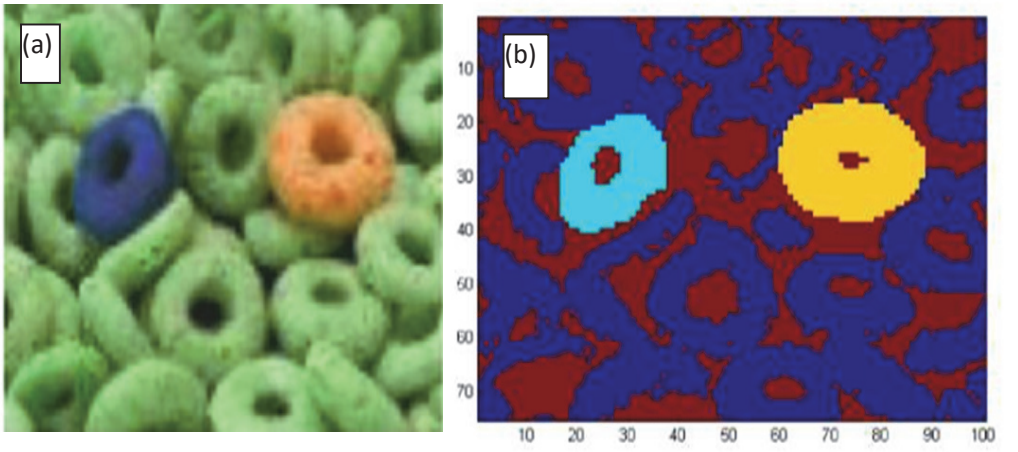

Figure 6. (a) Example for image segmentation application, **(b)** image segmentation applied with k largest eigenvalues

3.1 Projected Data Sets

We can use an image similar to the image in Figure (6-a) for segmentation application. Therefore the Distance matrix will be:

$$D(i,j) = \left\|\chi_i - \chi_j\right\|^2. \tag{3}$$

then the affinity matrix will be defined as:

$$e^{\frac{-D(i,j)^2}{\sigma^2}} \tag{4}$$

If the image size is w × h, (width × height per color plane in an RGB image Figure(6-b) then we shall have an affinity matrix of (w times h) squared.

So, even the image is practically small, say 100 pixels by 100 pixels, the size of the affinity matrix will be 10000 × 10000 which is quite large to be handled. Accordingly, we can employ the k largest eigenvalues to obtain an output And try to go further with proximal clusters with different geometry as showing in Figures: 7-a and 7-b respectively .

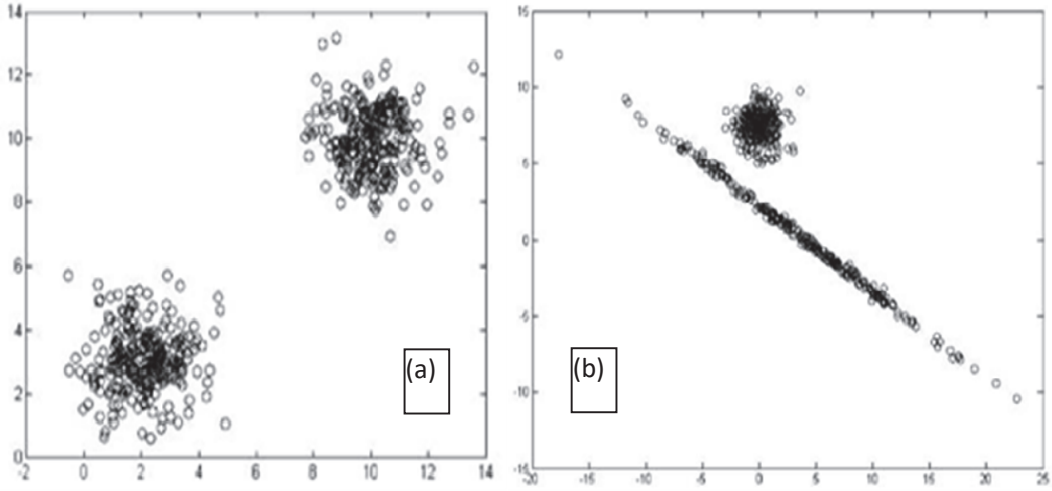

Figure 7. (a) Separated Clusters, **(b)** Proximal Clusters

3.2 Spectral Clustering Algorithm Steps [12]

- Given data points: $X_1, X_2, X_3, \ldots, X_n$
- Define similarities in pairs of data points;
  - $W_{ij} = s(X_i, X_j)$(Weighted adjacency matrix)
- Build a Similarity Graph G
  - Vertices (v_{ij}) are the data points
  - Edges $(e_{ij})$are the similarities
- Find Cut-throughs between sub-graphs
  - Define a cut objective function
  - Solve it

3.3 Implementation Steps

• Project the data into $R^{n\times n}$

• Represent the data in pairwise similarities

• Define a similarity/affinity matrix *W,* using a Gaussian Kernel $W_{ij} = e^{-\|s_i - s_j\|^2/2\sigma^2}$

• Set the diagonals $W(i, i) = 0$

• Define a diagonal matrix $D(i,j) = \sum_i W(i,j)$

• Form the normalized graph Laplacian by;

$$L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \quad (5)$$

• Compute the largest eigenvectors [16] of L; $x_1, x_2, x_3, \ldots x_k$ and place as columns into a new matrix X

• Form another matrix Y, by normalizing X

$$Y_{ij} = \frac{X_{ij}}{\left(\sum_j {X_{ij}}^2\right)^2} \epsilon R^{nxk} \quad 6$$

• Cluster the row vectors of *Y* by running K-means

• Note that:

- o x_i is assigned to a cluster only if row *i* of *Y* is assigned to that cluster.
- o Choosing *k*, and the scaling factor σ^2 are the big two challenges
- o Using K-means directly would not work with non-convex data sets.

3.3 Determination of Cluster Number

The most critical aspect of the clustering analysis is determining the number of clusters. However, many articles published today do not have definite conclusions on this issue. The most well-known equation from the first suggested approaches is as follows [17]:

$$k = \sqrt{\frac{n}{2}}. \quad (26)$$

Another method has been proposed by Marriott. It is therefore denoted by the letter M [18].

$$M = k^2 |W| \quad (27)$$

where W is the intra-group squares of the matrix and

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)\left(x_{ij} - \bar{x}_j\right)^t \quad (28)$$

where: n_j is the number of units in j^{th}, K is the cluster number, X_{ij} is the i^{th} unit values in the j^{th} cluster, and $\overline{X_j}$ is the sample mean vector of j^{th} cluster.

4. RESULTS

Firstly, we should consider that practically, the usage of cluster analysis is not for examining the data of a given technique. Cluster analysis is often employed to predict which variables combinations, similarity measures, and clustering techniques will lead to interesting and informative classifications. The analysis

proceeds in a few steps by interfering with the researcher to: change variables, select different similarity criterion, and to concentrate on a subset of a particular individual. Finally, the most important phase concerns the evaluation of the clustering solutions is obtained.

It should also be clear that any clustering method of that mentioned in the earlier sections can be determined to be the 'best' in each case. Certain methods are the best methods for some types of data. The table below shows some possible "methods of data combinations". It may be possible for many applications to implement a number of clustering methods. If all the solutions are mostly similar, the researcher may rightly be more confident than the results may be worth more research[4].

4.1 Experiments Screen Shots

The datasets reside in an excel workbook. Matlab application reads the first worksheet. We have 3 famous datasets so far and can adding more datasets if we like.

Table 2: dataset for Toy and real [19]

15.55	28.65	34.35	7.7	9.35	26.6	31.3	12.7	9.35	17.7	33.1	20.75	4.3	22.75	31.95	19.1
14.9	27.55	34.7	8	9.3	27.25	31.95	12.95	8.9	17.65	33.8	20.4	5.85	23.4	32.6	18.35
14.45	28.35	34.6	7.25	9.2	27.8	32.75	13.1	8.45	17.2	33.85	20	5.9	23.55	32.85	17.95
14.15	28.8	35	6.8	7.5	28.25	33.15	13.2	10.05	17.2	34.15	19.3	7.55	23.7	33.45	17.45
13.75	28.05	35.5	7.35	8.55	27.45	33.1	12.75	10.4	16.75	34.85	18.85	6.85	23.25	33.7	17
13.35	28.45	36.1	7.5	8.5	27.05	33.15	12.1	8.6	20.9	35.3	18.55	7.65	23.1	34.25	17.35
13	29.15	36.55	7	8.05	27.2	34.3	11.75	8.65	21.3	35.4	19.35	6.95	22.55	34.3	18.05
13.45	27.5	36	8.2	7.85	26.8	34	10.85	8.65	21.9	34.55	19.75	6.1	22.6	33.85	18.4
13.6	26.5	35.35	8.05	7.3	27.4	34.65	11	8.65	22.5	35.05	20	5.5	22.6	33.05	18.85
12.8	27.35	36.55	8.65	6.8	26.85	34.8	10.1	8.95	22.8	35.95	19.85	4.7	22.1	33.25	18.95
12.4	27.85	36.4	9.1	7	26.5	35.65	9.85	9.95	22.65	36.35	20.6	3.8	21.85	32.8	19.15
12.3	28.4	35.5	9.1	7.55	26.3	36.35	10	8.95	22.2	35.5	20.55	4.65	21.2	32.3	19.85
12.2	28.65	34.55	8.85	8.55	26.3	35.55	10.75	9.65	21.9	34.45	20.65	4.15	20.35	32.8	20
13.4	25.1	35.25	9.4	9	25.85	35.8	11.55	10.55	22.3	34.4	21.25	5.3	20.4	31.75	20.25
12.95	25.95	34.4	9.5	8.6	25.65	35.2	11.75	10.9	22.85	35	21.05	5.6	20.75	31.75	20.75
12.9	26.5	33.5	9.3	9.4	25.55	34.7	11.75	11.35	23.45	35.75	21.3	5.8	21.95	32.1	21.15
11.85	27	33.85	9.8	8.45	25.05	34.95	12.75	12.05	23.4	35.05	21.5	6.4	21.95	31.55	21.6
11.35	28	32.5	9.65	8.85	24.6	34.05	12.55	12.3	22.75	34.6	22.05	6.55	21.15	30.65	21.3
11.15	28.7	32.3	10.25	9.65	24.7	34.05	13.05	11.7	22.15	34.2	21.75	7.45	21.95	29.95	21.6
11.25	27.4	33.3	10.3	10.55	24.35	33.25	13.7	11.15	22.05	36.25	21.95	7.4	21.55	29.5	21.6
10.75	27.7	31.6	10.5	11.05	23.9	33.2	14.15	10.85	21.5	35.7	22.3	7.75	21.2	30.35	22.05
10.5	28.35	30.6	10.5	10.55	23.55	33.25	14.7	10.85	21.05	35.5	22.9	7.65	20.65	31.05	22
9.65	28.45	30.4	11.1	9.45	23.35	33	15.15	9.6	21.3	35.85	23.25	6.95	19.8	31.55	22.2
10.25	27.25	30.9	11.45	9.2	23.9	32.95	15.65	9.85	20.7	36.3	23.8	6.6	20.1	30.95	22.65
10.75	26.55	30.7	11.65	8.35	23.9	32.6	16.15	9.35	20.6	35.45	24.1	6.05	20.2	30.3	23.1
11.7	26.35	30.4	12.05	7.35	24.75	32.45	16.75	9.25	19.65	34.9	23.5	5.4	19.65	29.6	23.15

11.6	25.9	31.2	12	7.4	25.45	32.65	17.05	9.95	19.8	34.2	22.9	5.35	19.05	29.35	22.55
11.9	25.05	31.95	11.35	6.6	25.75	32.75	17.3	10.7	20.35	33.85	23.3	5.8	18.25	29.2	23.85
12.6	24.05	31.65	11.05	6.1	26	31.75	17.2	11.3	20.7	33.25	23.35	6.3	19.1	30.75	24
11.9	24.5	32.95	11.15	5.8	26.95	31.7	17.65	12.35	21.6	32.45	23.7	7	18.9	30.95	24.15
11.1	25.2	32.65	11.7	5.65	25.8	31	17.5	13.1	21.3	33.6	23.9	7.15	17.9	31.45	23.7
10.55	25.15	32.25	12.25	5.3	26.1	31.15	17.9	12.85	20.75	34.25	23.95	7.35	18.2	31.95	23.15
10.05	25.95	32.05	12.25	6.4	25.4	30.45	18.05	12	20	34.25	24.1	8.2	20.05	32.55	22.05
11	19.85	35.4	24.7	5.4	25.25	30.05	18.8	30.7	9.15	34.05	3.5	8.3	19.45	32.6	22.55
10.35	19	35.15	25.3	5.35	24.7	30.55	18.8	29.7	9.9	33.05	3.85	8.3	18.5	33.25	22.25
9.9	18.65	34.4	24.9	4.8	25.05	30.5	19.3	30.45	9.95	32	3.8	8.75	18.8	33.65	21.9
10.6	18.15	33.7	24.85	4.2	25.55	30.25	19.4	30.95	9.85	31.9	4.4	9.05	18.2	33.5	21.3
11.4	18.3	32.25	24.45	6.4	24.8	29.6	19.85	31.8	9.45	31.05	4.75	7.35	10.5	21	22.6
11.4	19.25	32.5	24.7	6.55	24.3	29.15	20.55	32.45	8.8	30.4	5.65	7.65	11.1	21.15	22.95
12.35	18.8	31.45	24.45	7.4	24.25	30.25	20.45	33.55	8.6	30.75	6.1	8.1	11.2	20.5	22.85
12.8	19.75	31.55	25.2	5.45	24.2	30.7	20.05	32.8	6	30	6.7	8.8	11.4	19.75	23.65
4.3	24	31	19.9	12.15	18.1	31.05	25	32.05	5.1	30.1	7.4	8.3	10.55	19.2	23.7
4	24.25	31.55	19.65	11.05	17.5	30.25	24.3	32.8	4.8	29.5	8.15	9	10.9	18.45	24.35
3.35	23.3	31.5	18.55	11.95	17.25	29.8	24.8	32.65	4.4	30.75	8	9.35	10.5	20.65	23.85
4.85	23.05	32.05	18.6	12.25	17.5	29.6	25.5	33.65	4.6	30.85	7.35	10.15	11	20.65	24.3
13.05	17.4	29.7	26.05	33.05	5.15	31.5	6.75	10.4	10.55	19.7	24.6	15.75	7.75	21.65	8.65
13.75	18.15	30.5	25.5	33.6	5.45	31.75	5.95	10.9	10	20.15	25.05	14.6	3.05	21.15	5.7
13.5	18.65	30.65	26	34.5	5.05	32.35	6.45	11.55	10.2	22.15	25.1	15	3.4	21.65	4.85
13.65	19.25	31.25	26.05	21.75	2.65	22.65	6.65	11.75	10.85	21.6	24.65	15.25	3.5	22.15	4.35
14	19.9	31.45	26.95	22	3.25	22.75	7.05	10.1	8.65	21.7	23.8	14.7	4.1	23.05	3.35
15.2	18.2	30.75	26.9	22.2	3.5	23	7.35	11.05	9.1	21.9	23.65	14.7	4.5	23.05	3.8
15.5	17.15	30.65	27.15	21.45	3.75	22.55	7.9	11.85	9.8	22.55	23.5	15.25	2.7	23.15	4.4
13.9	17.1	31.25	27.85	21.1	4.05	22.2	8.7	12.85	10.65	22.55	24.3	15.65	2.05	22.5	4.75
13.75	16.6	31.85	27.75	34.9	4.65	22.9	8.45	12.9	11.7	23.3	24.45	15.95	2.8	22.15	5.2
12.15	16.4	32.7	28.2	35.45	4.1	22.35	9.2	13.6	11.1	24.25	24.35	16.1	3.55	24.15	4.55
7.8	13.7	33.25	27.55	34.6	4.05	22.75	9.35	12.95	7.6	6.65	3.8	15.9	4	23.5	5.05
8.85	13.35	32.4	27.1	34.2	4.2	22.4	10.05	13.45	7.95	5.8	4	15.6	4.75	23.1	5.3
9	12.7	32.15	26.65	36.3	5.2	23.05	10.9	13.35	8.25	4.95	4.05	15.55	5.05	23	5.75
9.7	12.1	32.35	25.95	35.55	5.35	23.3	9.85	13.75	9	5.1	4.35	15.35	5.5	22.2	5.75
8.05	12.9	32.95	25.5	35.95	6.05	23.95	9.8	14.3	9.3	5.7	4.45	15.15	5.95	21.85	6.2
7.7	13.25	33.85	26.05	34.8	5.85	23.65	9.1	14.85	9.55	5.45	4.85	15.5	6.75	20.75	6.55
6.8	13.2	33.05	26.5	33.7	6.15	23.7	8.85	15.1	10.25	6.7	4.8	15.7	6.35	21	7.15
6.6	13.45	33.65	27	33.95	6.6	24.25	8.25	15.45	10.55	6.55	5.05	16.2	5.9	20.75	7.65
6.2	12.55	34.1	27.35	33.7	7.05	24.85	7.95	16.35	10.85	7.2	4.9	16.35	5.35	20	8.2
5.4	12.85	34.2	27.95	32.75	7.1	23.5	7.85	16.75	11.5	6.2	4.25	16.2	4.55	19.5	8.65
5.7	12.25	34.65	26.85	32.3	7.65	23.85	7.35	16.25	10.2	7.1	4.3	16.55	4.2	18.85	9.05
5.2	11.9	35.25	26	33	7.9	23.95	6.9	15.4	10.1	7.85	4.5	16.95	4.75	18.75	9.55

5.15	11.35	35.7	26.15	31.95	8.15	23.65	6.5	15.45	9.7	7.6	4.15	17.05	5.1	18.6	10
5.85	11.2	34.4	25.6	31.15	8.65	23.6	5.7	15.15	9.3	7.25	3.55	17.3	4.8	16.95	10.35
6.1	11.75	21.3	20.8	30.35	8.85	24.3	5.65	15.25	8.65	7.8	3.35	17.3	4.15	17.35	10.85
7	12.35	20.15	20.9	29.85	9	24.8	6.4	15.55	8.2	8.05	2.75	17.6	4.3	18	10.65
7.05	12.45	19.2	21.35	14.05	11.75	23.8	25.25	14.25	8.7	8.5	3.25	17.05	3.7	18.5	10.55
7.9	12.5	19.1	21.85	14.5	11.8	23.4	23.8	14.25	8.25	8.1	3.55	17.25	3.05	18.1	11.1
8.55	12.1	18.45	22.8	14.3	12.45	22.9	23.2	15.05	7.8	8.15	4	16.65	2.8	17.55	11.3
7.85	11.85	18.75	22.95	17	12.9	22.3	22.8	14.3	7.5	20.15	4.3	16.55	2.15	17.95	11.9
7.1	11.95	19.4	23	15.8	12.6	22.2	22.4	13.55	7.45	20.8	4.7	17.2	2.05	18.3	12
6.9	11.5	19.55	22.25	15.85	12	23.1	21.7	14.3	6.95	20.7	5.15	18.15	1.95	18	12.5
6.85	10.9	19.8	21.85	16.7	12.2	22.85	21.9	13.95	6.7	19.75	5.05	18.05	2.45	19	11.65
6.4	10.7	20.5	21.85	16.25	11.7	22.65	21.1	13.05	6.95	19.85	5.5	18.15	3.05	19.5	11.05
5.9	10.3	21.45	21.45	15.55	11.15	23.15	22.6	13.05	6.2	20.4	5.65	18.6	3.45	19.45	10.55
6.4	10.25	21.7	21.9	14.8	11.35	24.1	21.9	11.55	6.3	20.55	5.75	18.4	3.6	19.4	9.65
7.05	10.05	21.4	22.3	14.45	10.75	24.7	22.2	10.8	5.85	18.7	5.75	18.85	3.2	20.1	9.4
12.55	4.9	17.5	8.25	13.75	10.45	24.3	22.6	10.6	5.05	19.25	5.95	19.1	2.65	20.05	9.95
11.5	4.75	17.15	8.6	12.8	10.1	24.15	23.3	11.35	5.55	18.4	6	19.45	2.65	20.05	10.2
11.35	4.05	17.05	9	13.15	9.8	23.9	23.45	12.15	5.4	18.45	6.6	19	2.1	19.35	12.2
12.4	4.35	16.4	8.7	12.45	9.3	5.2	2.15	12.4	5.8	17.65	7.05	19.9	2.05	19.2	12.25
11.75	3.45	16.05	8.95	11.8	8.95	6.35	1.95	12.8	5.7	16.7	7.4	20.45	2.8	20.05	11.6
12.65	3.7	16.05	9.6	11.1	8.45	6.75	2.3	13.65	5.9	18.65	7.3	19.8	3.25	20.6	11.15
13.4	4.35	16.5	9.75	10.35	7.7	5.9	2.4	13.9	5.3	18.05	7.35	17.4	6.5	20.75	8.75
13.9	4.95	17.25	9.6	10.1	6.75	5.4	2.7	13.1	5.1	17.85	7.75	16.6	6.85	20.55	8.75
12.75	3	17.6	9.9	11.3	7.95	4.85	2.9	16.6	7.95	21.75	8.2	15.7	7.15	21.1	8
13.55	3.15	17.8	9.3	12.35	8.45	4.85	3.35	19.45	3.9	20.7	10.65	17.05	6.05	21.2	9.25
13.7	3.65	18	8.55	13.1	8.95	5.15	3.45	18.65	4.2	21.3	11.65	12.25	7.6	6.75	3.55
14.1	4.1	18.8	8.1	13.2	9.35	5.7	3.45	18.4	4.6	21.8	11.15	14.45	2.4	20.05	6.35
14.65	5.05	18.8	8.35	14.1	10.05	6.2	3	18.65	4.75	21.85	10.7	17.65	5.7	21.65	9.45
14.35	5.75	19.4	7.6	11.5	7.5	6.2	3.2	18.75	5.15	21.65	10.05	12.4	7.1	6.75	3.2
14.5	6.55	19.25	6.6	11.35	6.9	7.65	2.15	19.1	4.55	20.95	10.2	13.6	2.55	19.8	7.5
15.15	7.1	20.05	6.95	11.95	6.75	7.2	2.75	17.9	5.4	20.9	9.7				

4.2 Spectral Clustering by Matlab GUI

In geometrical, multi-range, and multi-level b GUI screen, the Spectral clustering controls is displayed at the leftmost side, we also applied the hierarchical and density based clustering outputs in the GUI to decide which is the best method for the specific dataset.

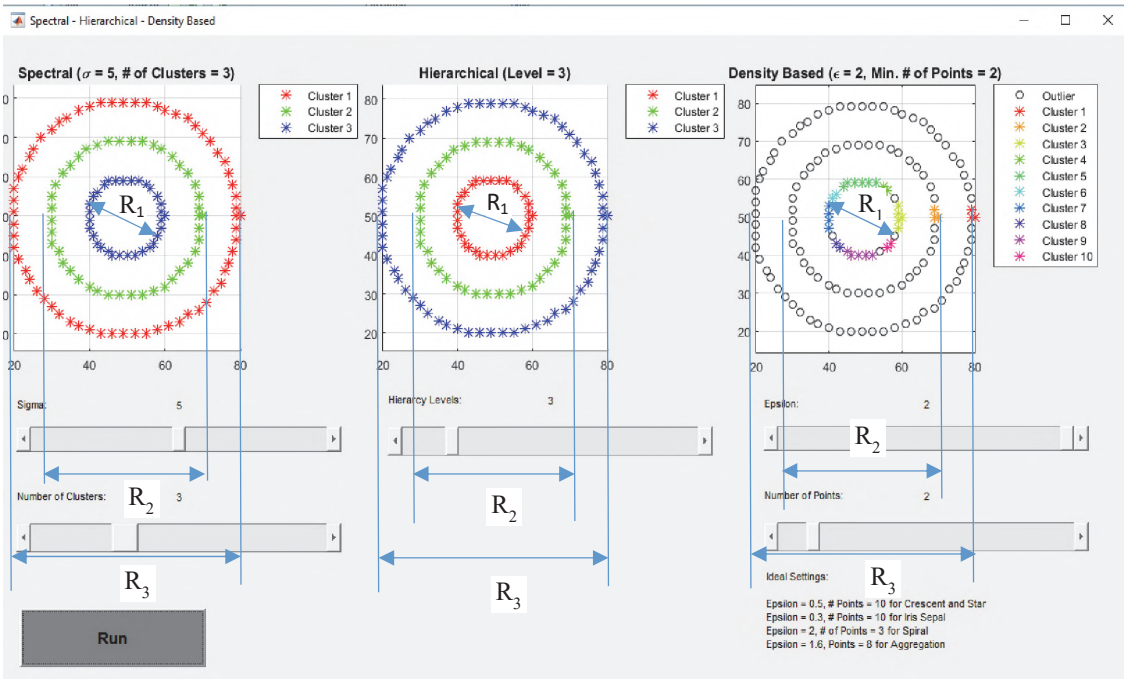

Figure 8. Concentric circles Spectral- Hierarchical-Density based

In Figure 8: the synthetic and the real datasets circles type are experimented by the three methods, which are, the spectral, hierarchical and density clustering methods. The parameters simulation are set as follows: the sigma and the number of clusters are equal to 5 and 3 respectively. Where the similar datasets of synthetic and real are 1000 pairs and have chosen in the range 20-80 for synthetic (x-axis) and the real datasets. In this figure, the clustering by spectral and hierarchical methods are better than the density method in case of the diameter is vary in all ranges R_1, R_2, and R_3 diameters. Where, R_1 range is equal to 20-40, R_2 = 30 -70, and R_3= 20- 80. Furthermore, at each R_i, $i \in \{1,2,3\}$, and for fixed diameter (i.e. R_i is constant), it is preferable to apply the clustering by density method where it produces a superior clustering in comparison with the spectral and hierarchical methods clustering.

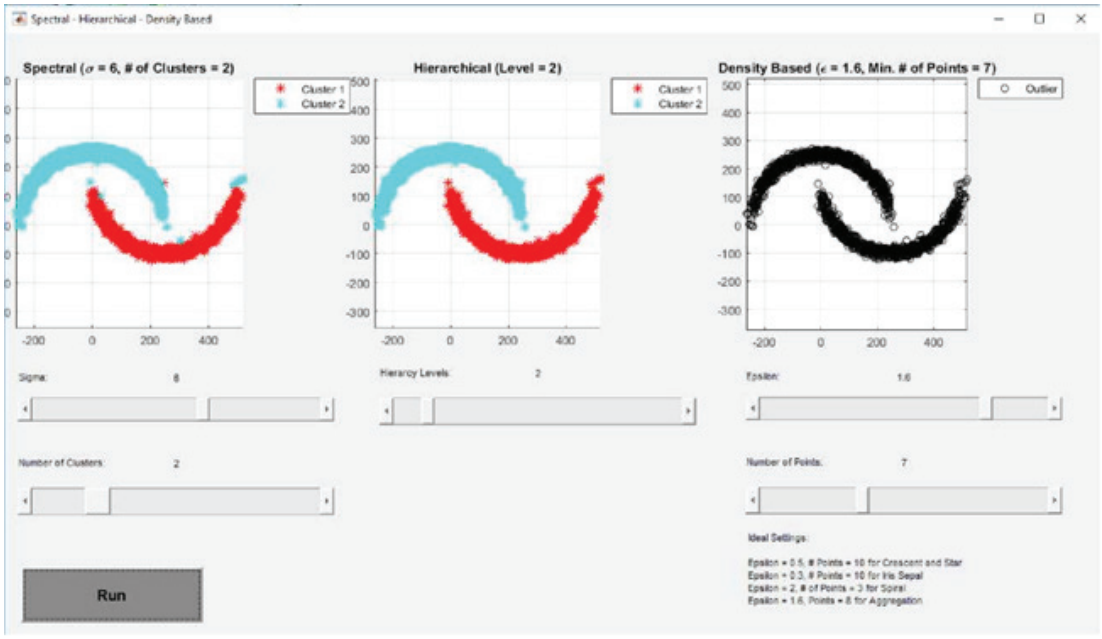

Figure 9. Semicircles: Spectral- Hierarchical-Density based

In Figure 9: the synthetic and the real datasets Semicircles type are experimented by the three methods, which are, the spectral, hierarchical and density methods. The parameters simulation are set as follows: the sigma and the number of clusters are equal to 5 and 2 respectively. Where the similar 2000 datasets of the synthetic and real are chosen in the range of -250 to 500 for synthetic dataset (x-axis), where the real dataset in the range of -150 to 300. In this figure, the clustering by spectral and hierarchical methods are better than the density method. Furthermore, the clustering by employing the spectral method produces superior clustering in comparison with the hierarchical method as well as the density method clustering too. This means, the spectral method is recommended to do clustering of dataset at upper values in the range -250 to 500.

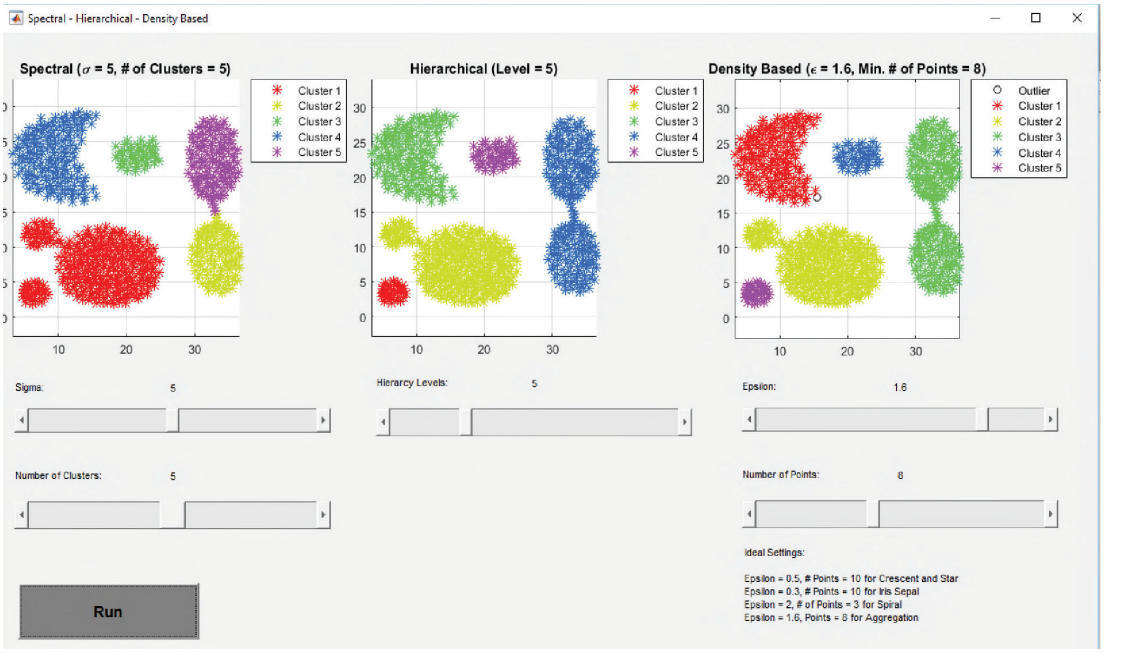

Figure 10. Aggregation: Spectral- Hierarchical-Density based

In Figure 10: Aggregation dataset type is experimented by the three methods, which are respectively, the spectral, hierarchical and density methods. The parameters simulation are set as follows: the sigma and the number of clusters are equal to 5. Where the similar datasets of synthetic and real are chosen within 788 values and in the range of 0-40 for synthetic dataset (x-axis), where the real dataset in the range of 0-30. In this figure, in the range of 30-40 for synthetic and values 20-30 of real dataset, the clustering by spectral method is better than the other methods (i.e. hierarchical and density methods), where at values that are under 10 for both synthetic and real datasets, the clustering by utilizing the hierarchical and density methods is better than the spectral method. This means, the spectral method is recommended to do clustering of dataset that has similar values upper than 20, where it is not recommended when the similar values are lower than 10.

5. CONCLUSIONS

Arrangement in huge databases is troublesome. Utilizing clustering calculations is good method for grouping which is based in this study. There are different methods for characterization and they have a few issues. It is the primary assignment to locate the right information parameters, to discover clusters of discretionary shapes and to make the entire procedure in a sensible time. The spectral algorithm solves all these problems because it can find complex shape sets very quickly. We discussed the usability of the spectral algorithm compared to the known hierarchical and density algorithms. Where spectral algorithm has superior clustering against the two other algorithms in big datasets, in some cases, in small datasets, the hierarchical and density algorithms have advantages in clustering from spectral algorithm.

6. REFERENCES

S. Sharma, "Applied Multivariate Techniques," *John Willey Sons*, 1996.

[**H. Tatlıdil,** "Uygulamalı çok değişkenli istatistiksel analiz," *Akad. Yayınları*, 1996.

Q. Li, Y. Ren, L. Li, and W. Liu, "Fuzzy based af fi nity learning for spectral clustering," *Pattern Recognit.*, vol. 60, pp. 531–542, 2016.

I. B. Society, "A General Coefficient of Similarity and Some of Its Properties Author (s): J . C . Gower Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2528823 International Biometric Society is collaborating with JSTOR to digitize , preserve and extend," vol. 27, no. 4, pp. 857–871, 2018.

[**F. H. C. Guttiérrez Toscano, P.,& Marriott,** "Unsupervised classification of chemical compounds. Journal of the Royal Statistical Society," *Ser. C (Applied Stat.*, vol. 48, no. 2, pp. 153–163, 1999.

C. Hair Jr, J. F., Anderson, R. E., Tatham, R. L., & William, "Multivariate data analysis with readings.," *New Jersy Prentice Hall.*, 1995.

M. R. Anderberg, "Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks.," *Acad. Press*, vol. 19, 2014.

M. S. Blashfield, R. K.,& Aldenderfer, "The literature on cluster analysis. Multivariate Behavioral Research," vol. 13, no. 3, pp. 271–295, 1978.

S. H. Spielmat, D. A.,& Teng, "Spectral partitioning works: Planar graphs and finite element meshes.," *Found. Comput. Sci. Proceedings., 37th Annu. Symp. IEEE.*, pp. 96–105, 1996.

M. Everitt, B. S., Landau, S., & Leese, "Clustering analysis," *Arnold, London*, 2001.

M. Han, J., Pei, J., & Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

A. Ş. Çelik, M., Dadaşer-Çelik, F., & Dokuz, "Anomaly detection in temperature data using dbscan algorithm.," *Innov. Intell. Syst. Appl. (INISTA), 2011 Int. Symp. IEEE*, pp. 91–95.

M. Ester, H. Kriegel, X. Xu, and D.- Miinchen, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996.

H. (Eds.). Ho, T. B., Cheung, D., & Liu, "Advances in Knowledge Discovery and Data Mining," in *9th Pacific-Asia Conference, PAKDD 2005, Hanoi, Vietnam, May 18-20, 2005, Proceedings(Vol. 3518). Springer.*, 2005.

G. Yazgan, E.,& KAYAALP, "Kümeleme (Cluster) Analizi Yöntemlerinin Karşılaştırmalı Olarak İncelenmesi ve Tarımsal Araştırmalarda Kullanılması.," Zootekni Anabilim Dalı, Adana., 2002.

R. Atkinson, Q., Nicholls, G., Welch, D., & Gray, "From words to dates: water into wine, mathemagic or phylogenetic inference?," *Trans. Philol. Soc.*, vol. 103, no. 2, pp. 193–219, 2005.

A. Kannan, R., Vempala, S., & Vetta, "On clusterings: Good, bad and spectral," *J. ACM (JACM), 51(3), 497-515.*, 2004.

F. H. C. Marriott, "Practical Problems in a Method of Cluster Analysis," *Biometrics*, vol. 27, no. 3, pp. 501–514, 1971.

https://www.mathworks.com/matlabcentral/fileexchange/34412-fast-and-efficient-spectral-clustering