Pattern Recognition

Project Report

Clustering by fast search and find of density peaks

Implementation :

1) Loading data set: We have taken 3 datasets and asked user to choose one
   a) Crescent moon random dataset
   b) Dataset given in IEEE paper
   c) Flame dataset
2) Calculating the distance matrix ,we have used eucledian distance
3) Calculate cut off distance
   Dc(cut off distance) depends up on Average number of neighbours is around 1 to 2 % of total number of points in the datasets
4) Calculating local density Rho
5) Calculating delta for each point
   delta is measured by computing the min distance between the point I and any other point with higher density
6) Based on rho and delta values we assign points to the cluster or as an outlier

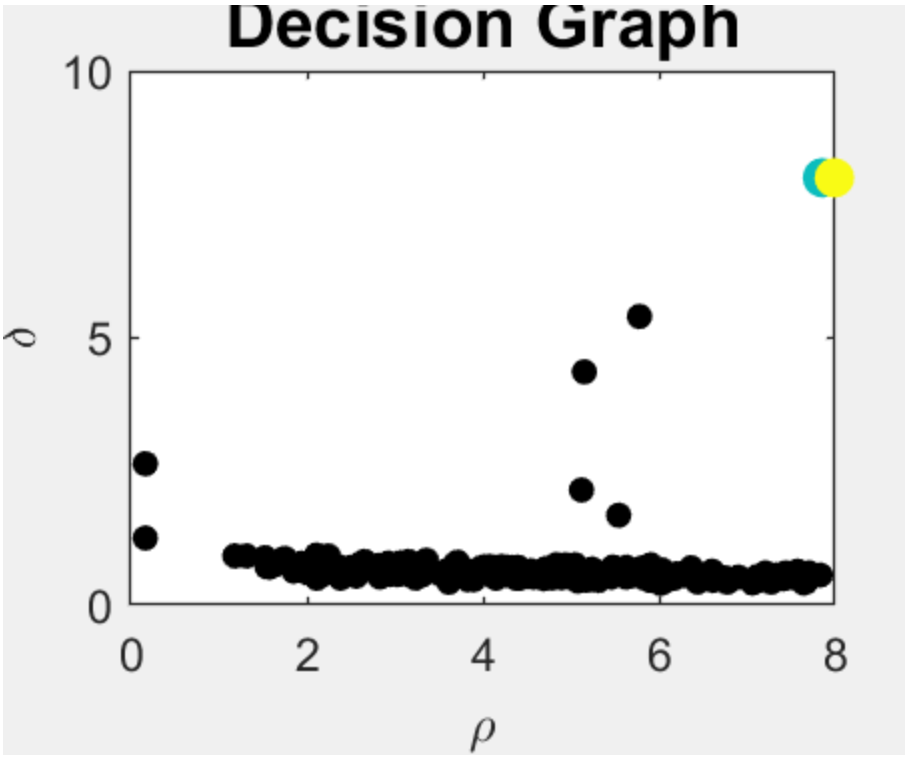
OUTPUTS

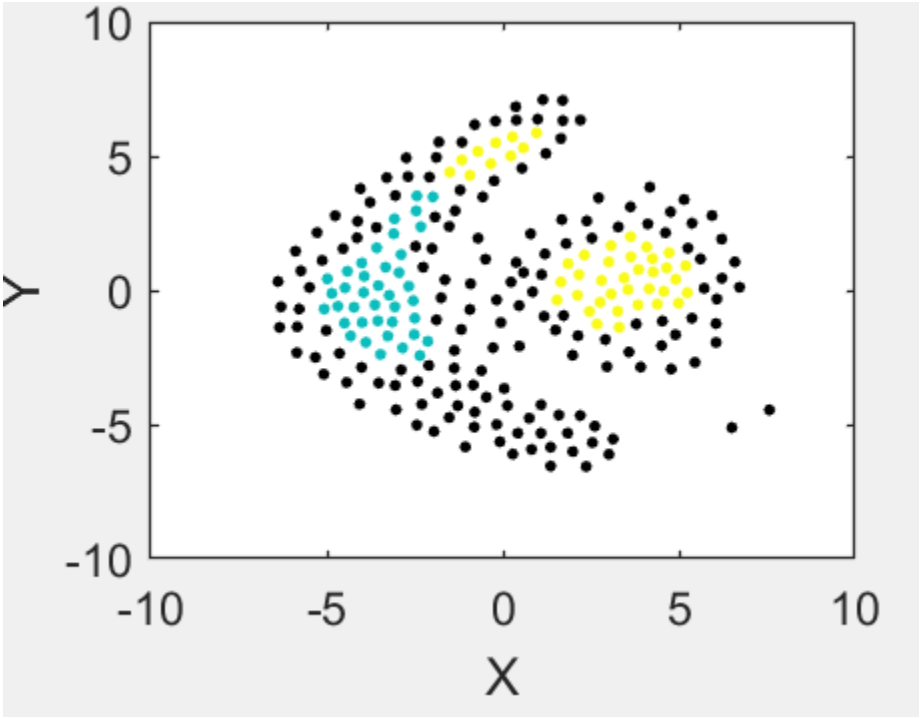
DataSet: Flame.txt

Enter 1 for flame dataset

No of clusters: 2(pick from decision graph)

No of points:240

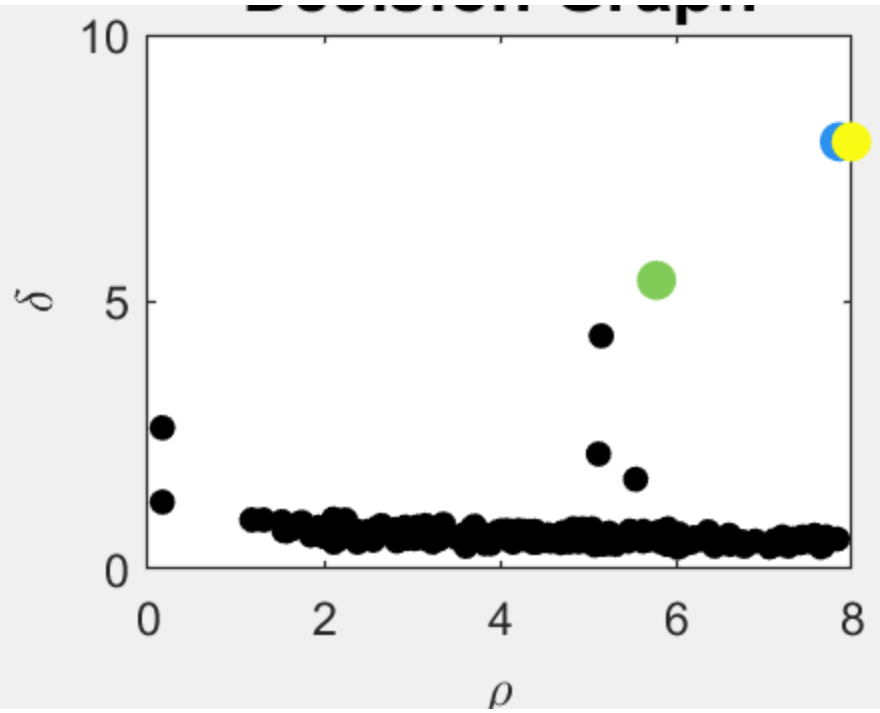


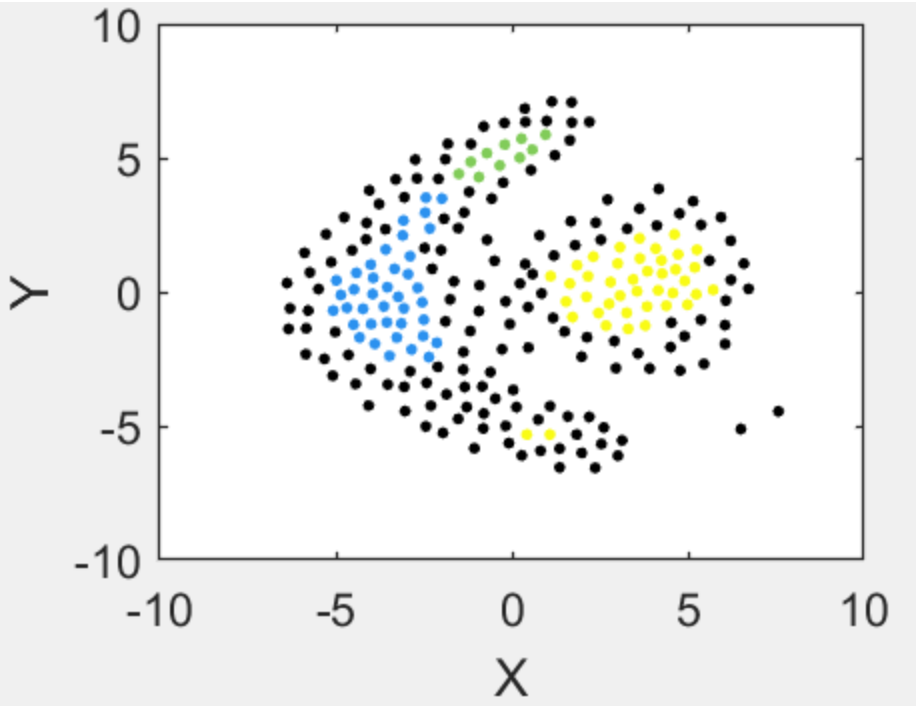


CLUSTER: 1 CENTER: 69 ELEMENTS: 102 CORE: 39 HALO: 63

CLUSTER: 2 CENTER: 230 ELEMENTS: 138 CORE: 43 HALO: 95

Clusters: 3 (pick from decision graph)





CLUSTER: 1 CENTER: 69 ELEMENTS: 102 CORE: 39 HALO: 63

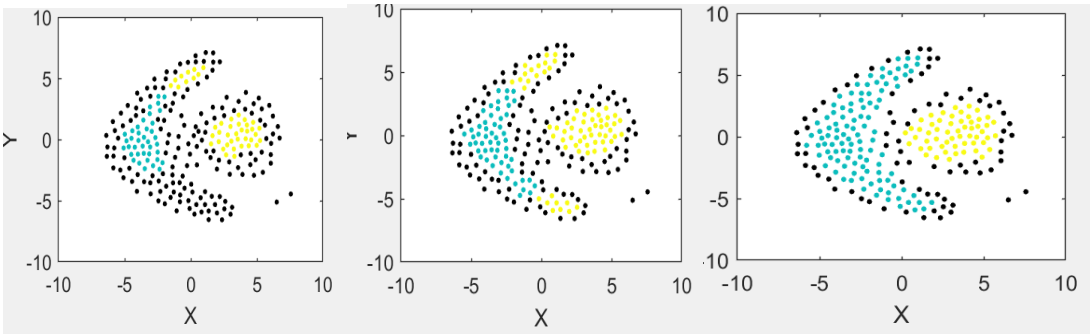
CLUSTER: 2 CENTER: 138 ELEMENTS: 32 CORE: 10 HALO: 22

CLUSTER: 3 CENTER: 230 ELEMENTS: 106 CORE: 41 HALO: 65

Experiment1:changing percentage of average number of neighbours

no of clusters=2 we got the following result

Percent=1 percent=2 percent=4



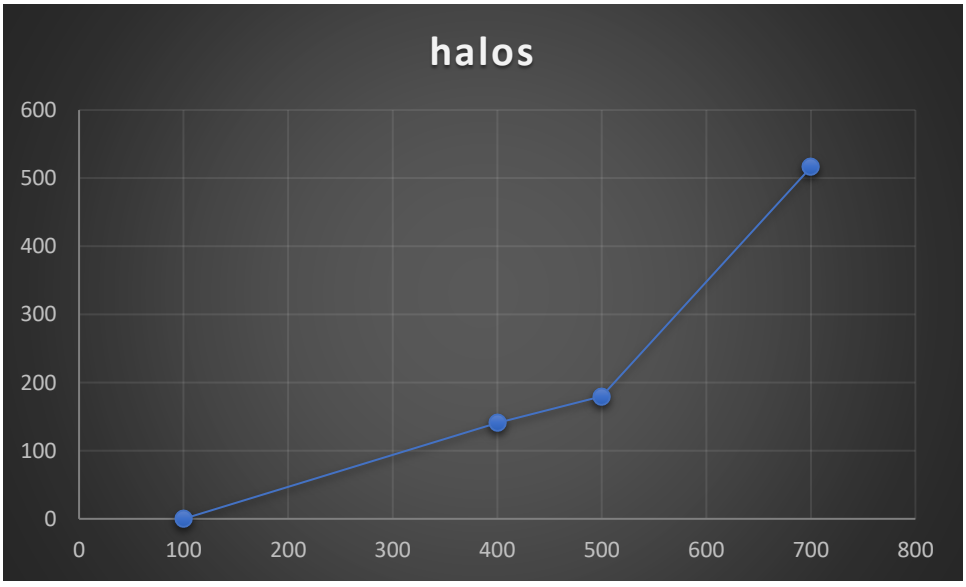
We observed for percentage of avg, number of neighbours is 4 we got better clustering.

Experiment2:

Number of Halos(outliers) Vs number of Samples

X axis: number of samples

Y-axis: number of outliers



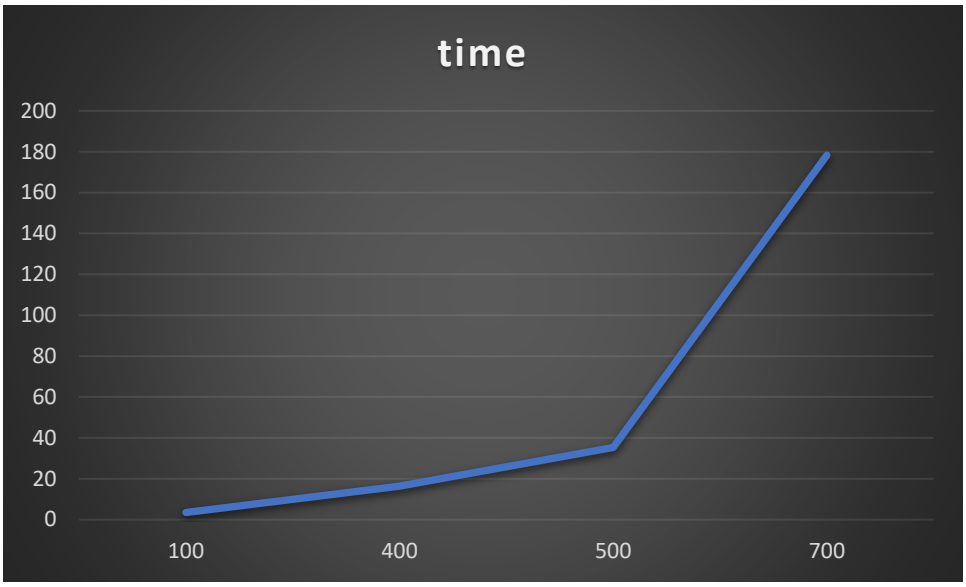
As number of samples increases halos increased.

Experiment3:

ExecutionTime vs number of samples:

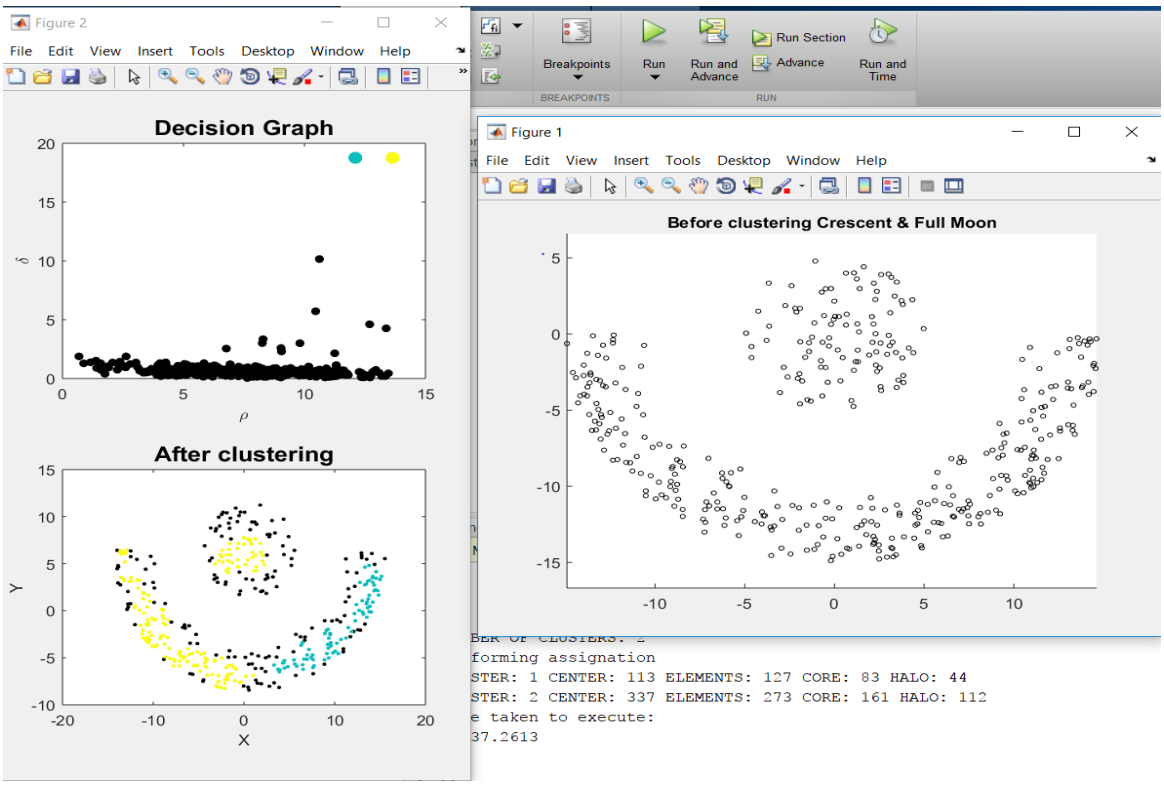
X-axis: number of sample

Y-axis: time in secs



As number of points increases execution time increases.

Experiment4:



enter 1 for Random CrescentFullMoon data 2 for data set present in IEEE paper 3 for Flame data set :1

159600 3

Generated file:DECISION GRAPH

column 1:Density

column 2:Delta

Select a rectangle region to pick cut off values of rho and delta for cluster centers

ans =

0.3000

NUMBER OF CLUSTERS: 2

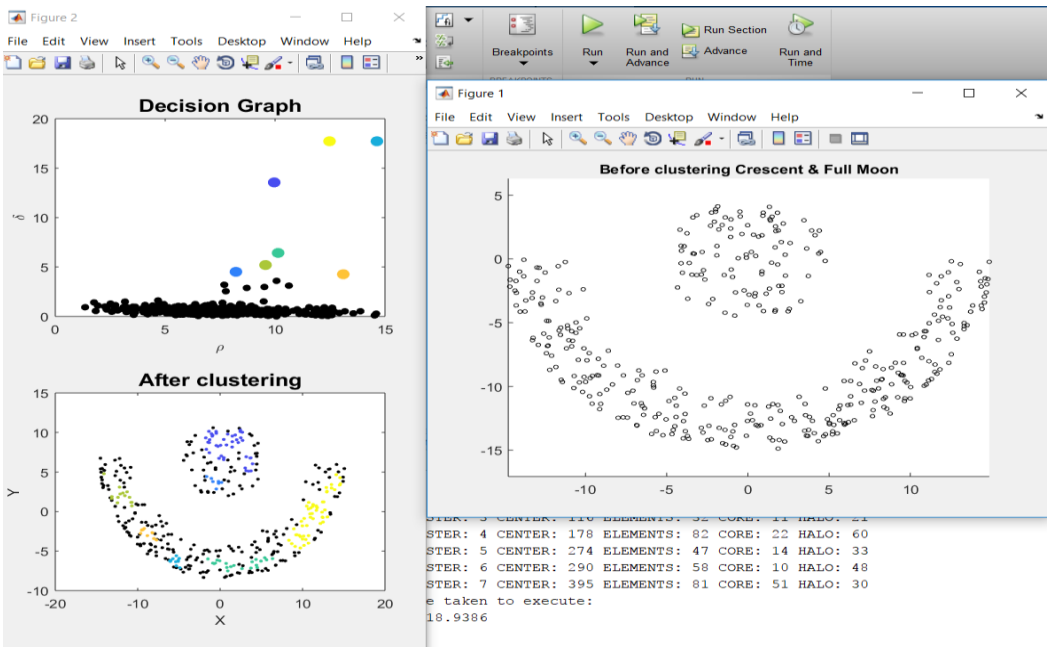
Performing assignation

CLUSTER: 1 CENTER: 113 ELEMENTS: 127 CORE: 83 HALO: 44

CLUSTER: 2 CENTER: 337 ELEMENTS: 273 CORE: 161 HALO: 112

time taken to execute:

37.2613



NUMBER OF CLUSTERS: 7

Performing assignation

CLUSTER: 1 CENTER: 2 ELEMENTS: 72 CORE: 41 HALO: 31

CLUSTER: 2 CENTER: 6 ELEMENTS: 28 CORE: 7 HALO: 21

CLUSTER: 3 CENTER: 116 ELEMENTS: 32 CORE: 11 HALO: 21

CLUSTER: 4 CENTER: 178 ELEMENTS: 82 CORE: 22 HALO: 60

CLUSTER: 5 CENTER: 274 ELEMENTS: 47 CORE: 14 HALO: 33

CLUSTER: 6 CENTER: 290 ELEMENTS: 58 CORE: 10 HALO: 48

CLUSTER: 7 CENTER: 395 ELEMENTS: 81 CORE: 51 HALO: 30

time taken to execute:

18.9386

Advantages:

1) We can select choose number of clusters dynamically
2) It is robust to number of instances

Limitations:

1) We have taken 2 dimensions data only
2) Selecting rectangular region is for cut off rho and delta values
3) Varying cutoff distance produce consistent results if it is 1-2%

Drawbacks:

1) The paper doesnot talk about noise
2) As number of samples increases performances reduces

References:

https://www.mathworks.com/matlabcentral/fileexchange/41459-6-functions-for-generatingartificial-datasets

http://cs.uef.fi/sipu/datasets/

https://www.mathworks.com/matlabcentral/fileexchange/53922-densityclust