TO FIND THE NATURAL GROUPING PRESENT IN A GIVEN DATA SET USING MST OF DATA POINTS

Objective of the work

- *To find the "natural grouping" of a given data set using MST of data points.*

- *Clustering techniques aim to extract such "natural groups"*

  present in a given data set and each such group is termed

  as a cluster.

CLUSTERING

Let the set of n patterns $S=\{x_1, x_2, \ldots, x_n\} \in \Re^m$ *and* K *clusters are represented by* $C_1, C_2, \ldots, C_K$ *then*

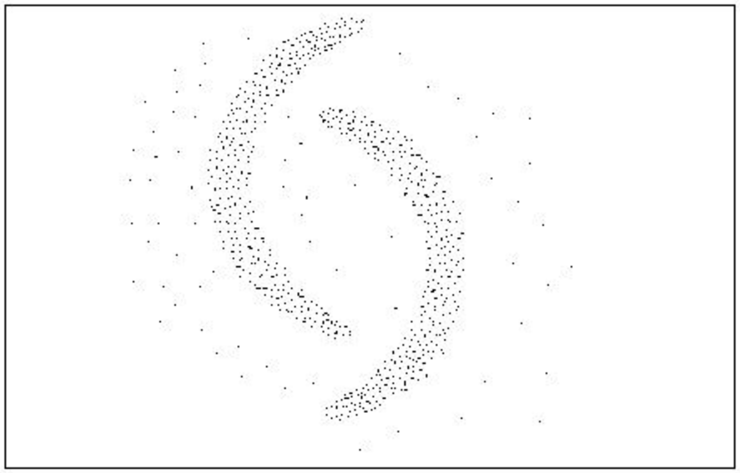
1. $C_i \neq \varphi$, *for* $i= 1, 2, \ldots, K$
2. $C_i \cap C_j = \varphi$ *for* $i \neq j$ *and*
3. $\cup_{i=1}^{K} C_i = S$ *where* φ *represents null set.*

What is Natural Grouping?

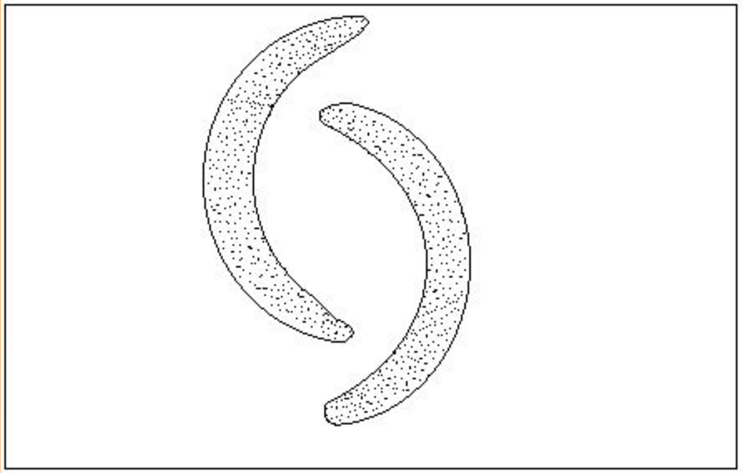
- *For a data set* $S = \{x_1, x_2, \ldots, x_n\} \in \Re^m$ *, what one perceives to be the groups present in* S *by viewing the scatter diagram*

  of S, is termed as natural groups of S.

Natural Grouping

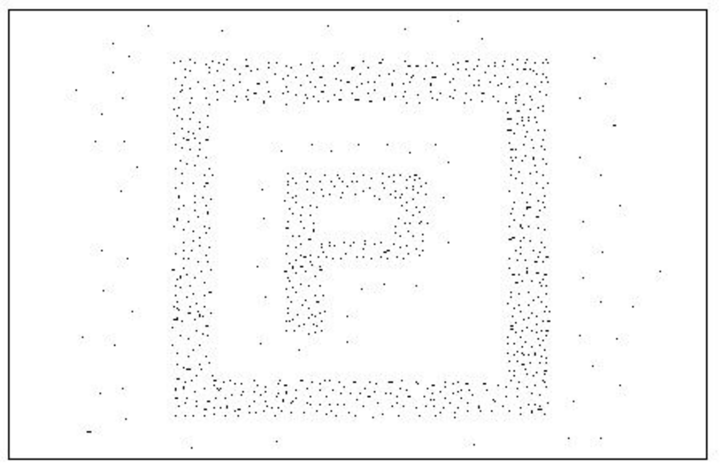


Scatter Diagram

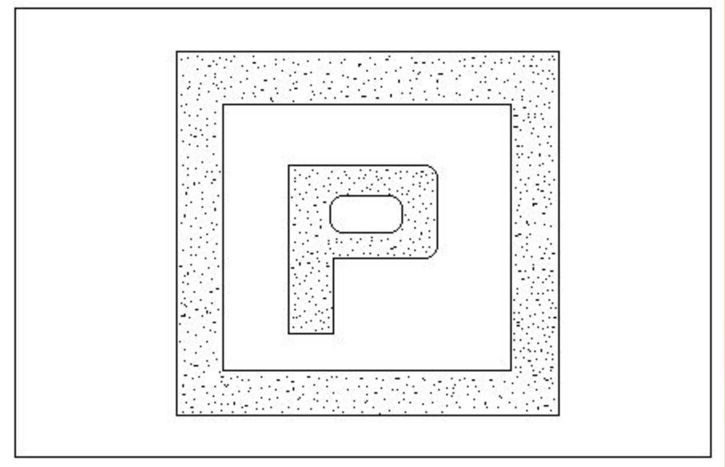


Natural Grouping

Natural Grouping



Scatter Diagram



Natural Grouping

Widely used Algorithms

- *K-Means Algorithm.*
- *ISODATA Algorithm.*

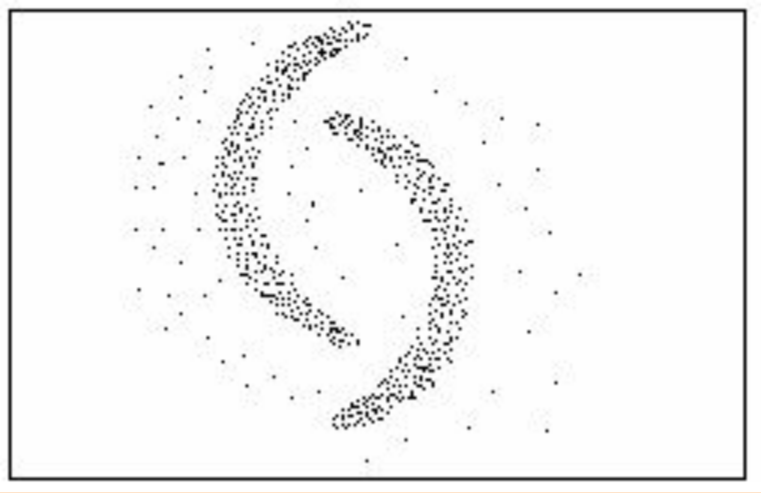
Disadvantage of K-Means Algorithm

- *It needs the number of clusters to be known a priori.*
- *It can find the grouping if clusters exhibit characteristics pocket and are not placed very close to each other.*
- *It may stack at a local minima.*
- *It can not provide proper grouping in case of some data having typical shape and size.*

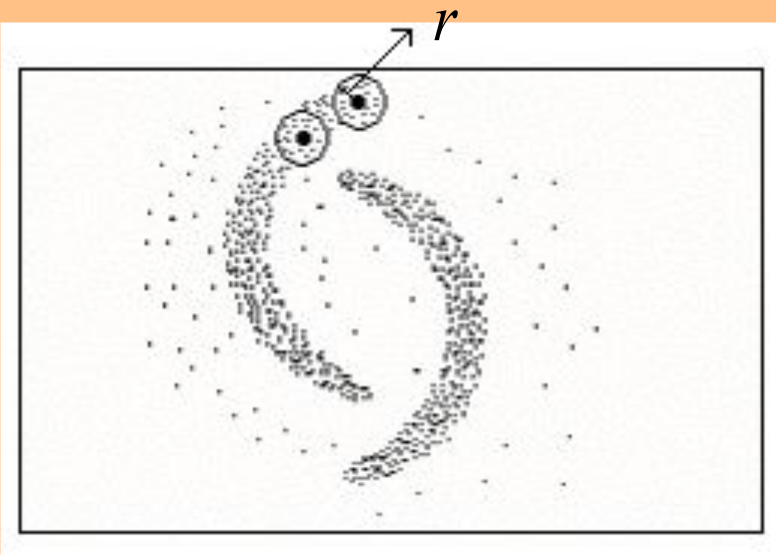
To find the natural grouping based on local densities of the data points

- *To find high density regions of a given data set.*
- *To merge those high density regions "suitably" along with the data*

  points of other regions to result in clustering.
- *To eliminates noise if any from the final clustering.*



Scatter Diagram

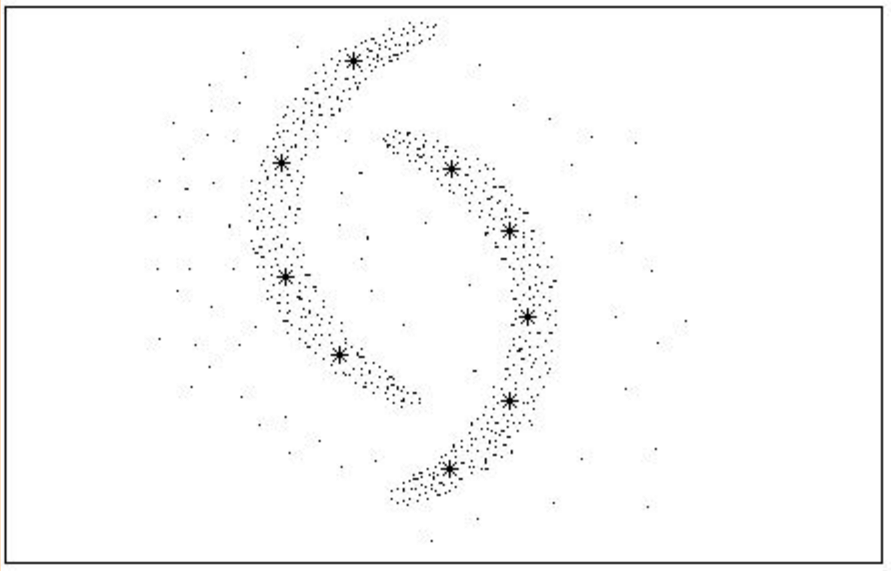


Finding density of each data points

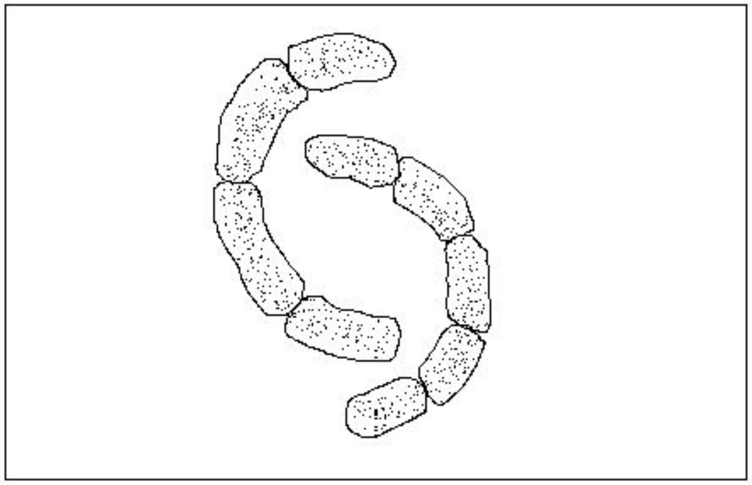
Radius for the open disk to compute the density of each data point is taken to be

$$r = h_n = \left(\frac{l_n}{n} \right)$$

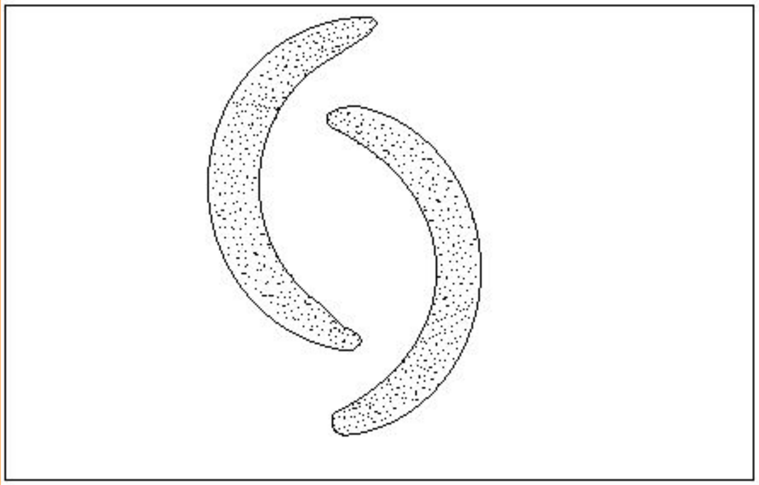
l_n → Sum of edge weights of minimal spanning tree of S.
n → Number of data points in S.
The edge weight is taken to be the Euclidean distance.



Scatter diagram with seed points



Scatter Diagram with 9 Groups



Natural Grouping By Proposed Method

Algorithm AL-1:

Step 1: Let $S=\{x_1,x_2,\ldots\ldots,x_n\}\subseteq\Re^m\ (m\ge 2)$. *Find the MST of* S *with the edge weight as the Euclidean distance. Let* $h_n=\left(\frac{l_n}{n}\right)$ *and the radius* $r=h_n$.

Step 2: Compute the density (the number of points) for each datum x as the number of other data units within an open disc of radius h_n *with x as center.* m_i *denote the density of the point* x_i, *i=1, …, n. In other words, let*

$$A_i=\{y:\|x_i-y\|\le h_n,\ y\in S\},\ i=1,2,\ldots,n$$

and

$$m_i=\#A_i,\ i=1,2,\ldots,n.$$

(#A means the number of points of the set A).

Step 3: Rearrange $m_1, m_2, \ldots\ldots, m_n$ in increasing order. Let the rearrangement be $m_1^*, m_2^*, \ldots\ldots, m_n^*$. Let $p_j, j = 1, 2, \ldots., n$ represent the corresponding cumulative sums of $m_1^*, m_2^*, \ldots\ldots, m_n^*$. i.e.

$$p_j = \sum_{i=1}^{j} m_i^*, j = 1, 2, \ldots.., n.$$

Step 4: Compute $M = \left[\frac{w}{100} \times n \right]$ where [a] means integral part of 'a'

i.e. the largest integer ≤ a. Find the value of i for which p_i is nearest to M.

Choose $k = m_i^*$ for that i.

If $p_i < M < p_{i+1}$ and $M - p_i = p_{i+1} - M$ then choose $k = m_{i+1}^*$.

We have taken the value of w to be 85.

Step 5: *Find the set $S_1 \subseteq S$ such that every point in S_1 has density at least equal to k i.e. Let* $S_1 = \{x_i : m_i \geq k,\ x_i \in S\} \subseteq S$. S_1 *represents the set of high density points of S.*

Step 6: *Arrange the points of S_1 according to their density. Choose the point having highest density as the first seed point.*

Step 7: *Choose subsequent seed points from the array of points from S_1 subject to the stipulation that each new seed point is at least at a distance $2h_n$ from all other previously chosen seed points. Continue until all remaining data units of S_1 are exhausted. Let V be the set of seed points of S. Let t = #V. Let V = { z_j , j=1, 2, …,t }. V is the*

Step 8: *Stop.*

ALGORITHM AL-2

Step 1: *Let* $V = \{z_j, j = 1, 2, ..., t\}$ *be the set of seed points of*

$$S = \{x_1, x_2, ..., x_n\} \subseteq \Re^m (m \geq 2).$$

Step 2: *Divide the n points of S into t groups in the following way:*

Put x_i *in the* j_{th} *group* C_j *utilizing the minimum squared Euclidean distance classifier concept: i. e.* $x_i \in z_j$ *if*

$$\left\| x_i - z_j \right\|^2 < \left\| x_i - z_q \right\|^2 \quad \forall q \in \{1, 2, ..., t\},\ q \neq j.$$

and $\quad \cup_{q=1}^{t} C_q = S.$

Step 3: *For two groups C_i and C_j find d_{ij} where* $d_{ij} = Min\{d(x_{m1}, x_{m2})\}$, $x_{m1} \in C_i, x_{m2} \in C_j$. *If* $d_{ij} \le h_n$ *then merge those two groups C_i and C_j into one group and name it as C_i.*

Step 4: *Repeat Step 3 for all possible pairs of i and j.*

Step 5: *Stop.*

Algorithm to Eliminate Noise

Step 1: Let $C_i, \ \{i = 1, 2, ..., K\}$ be the K clusters of

$$S = \{x_1, x_2, ..., x_n\} \subseteq \Re^m (m \geq 2).$$

Step 2: For each group C_i, compute the distance $d_j, \forall x_j \in C_i$

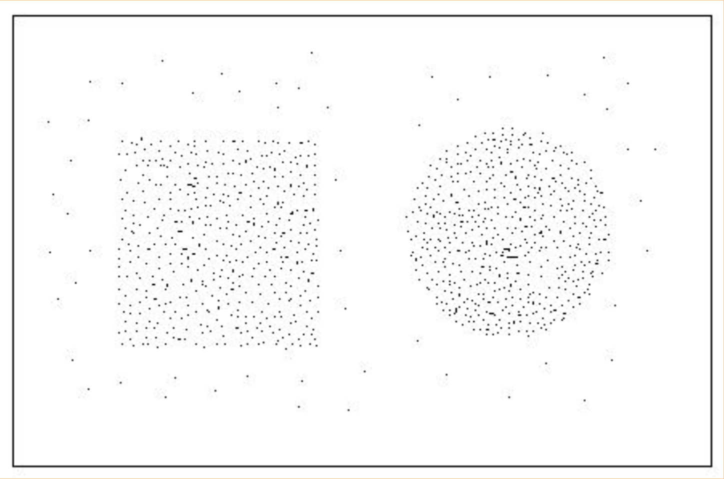
where $d_j = Min\{d(x_j, x_l)\}, \ j \neq l, \ x_l \in C_i$. If $d_j > 2h_n$

then remove the data point x_j from the clustering obtained by AL-2.

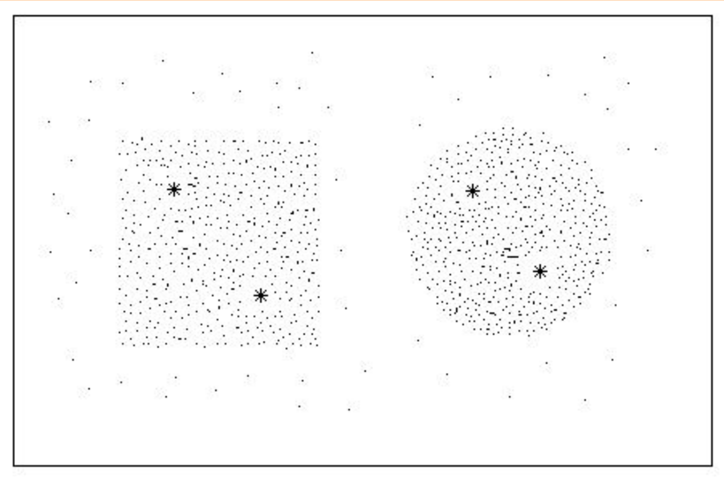
Step 3: Repeat Step 2 for all possible pairs of i and j.

Step 4: Stop.

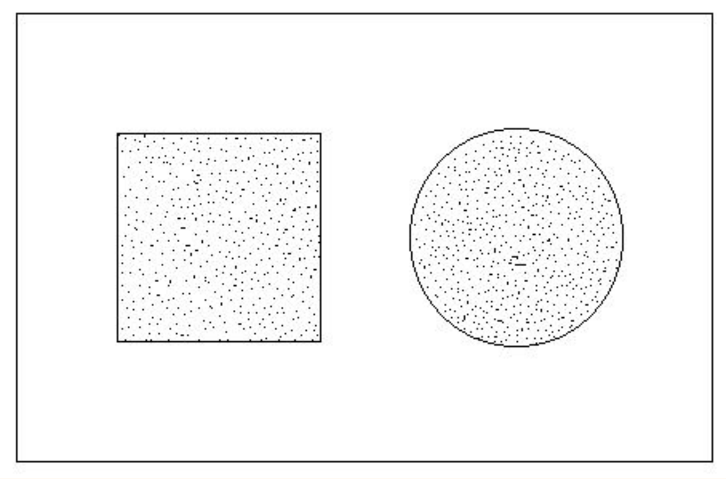
Experimental Results



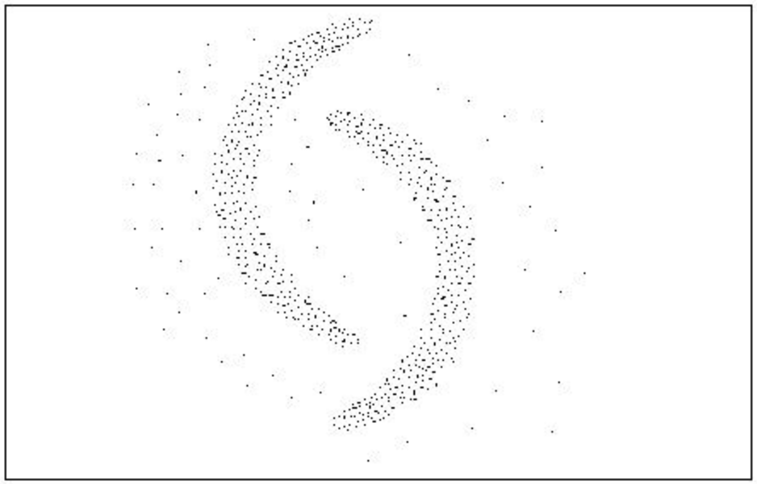
Scatter diagram of synthetic data with noise



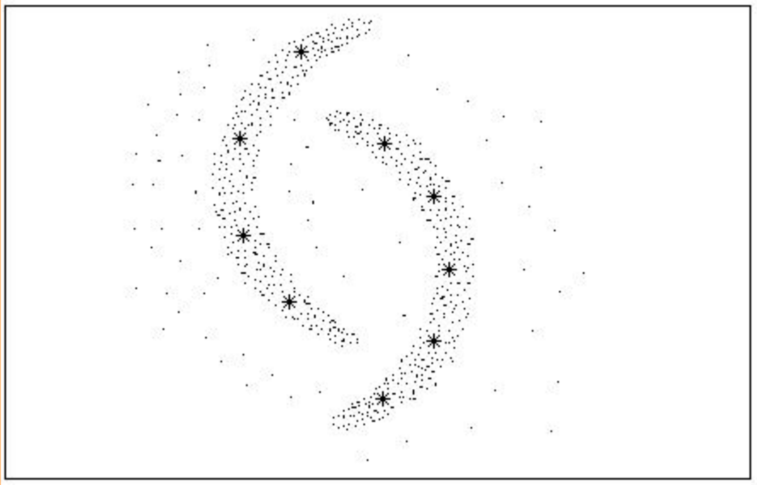
Scatter diagram with seed points



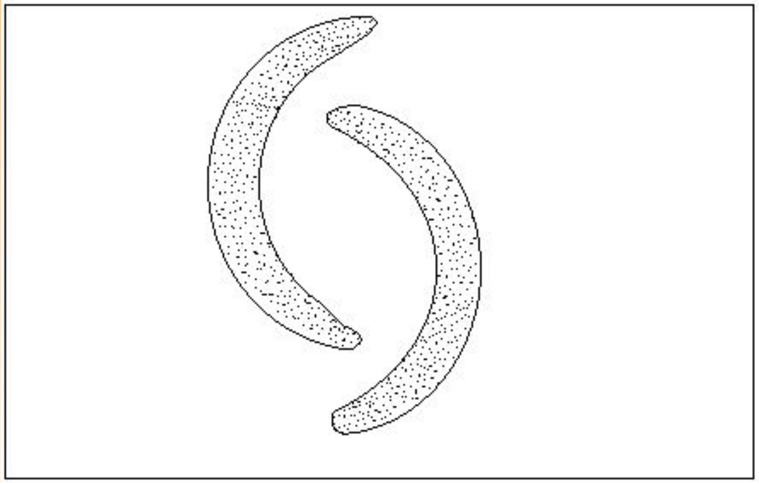
Clustering by the proposed method



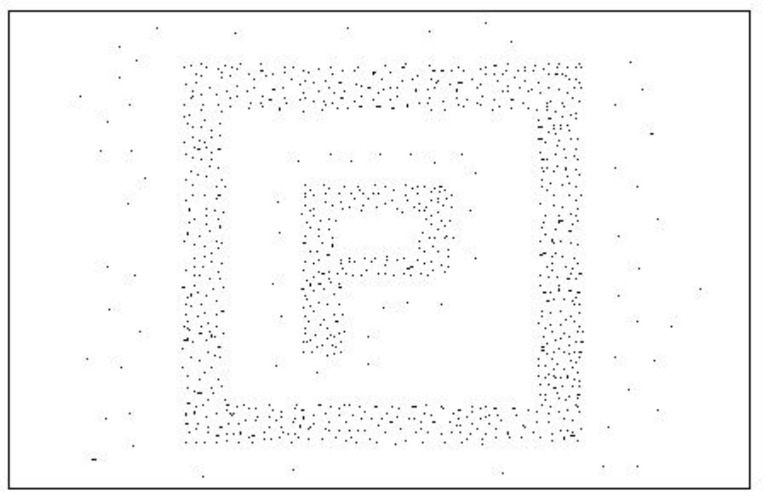
Scatter diagram of synthetic data with noise



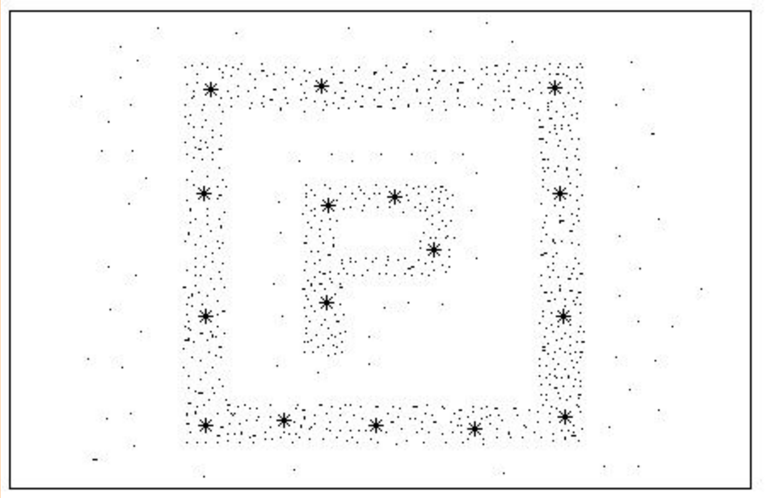
Scatter diagram with seed points



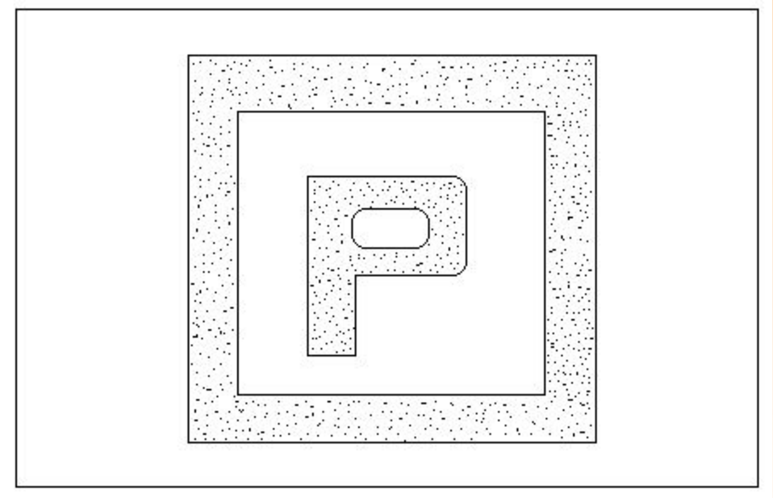
Clustering by the proposed method



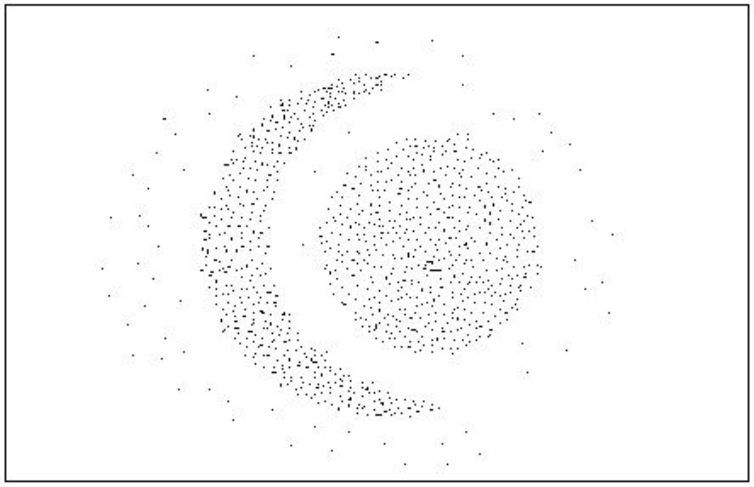
Scatter diagram of synthetic data with noise



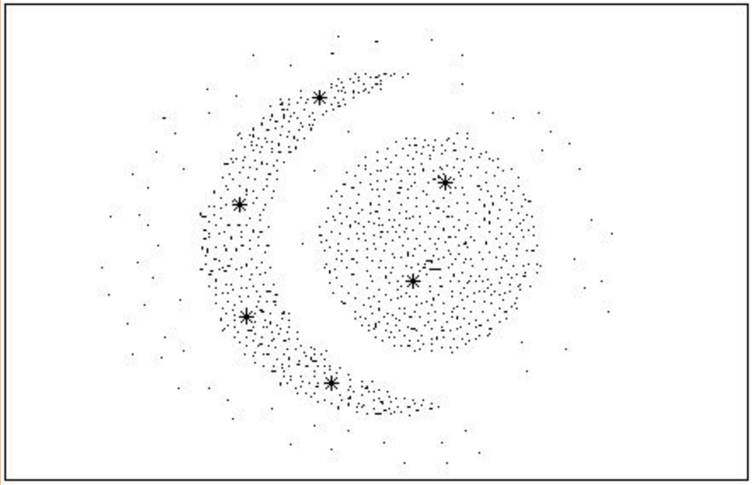
Scatter diagram with seed points



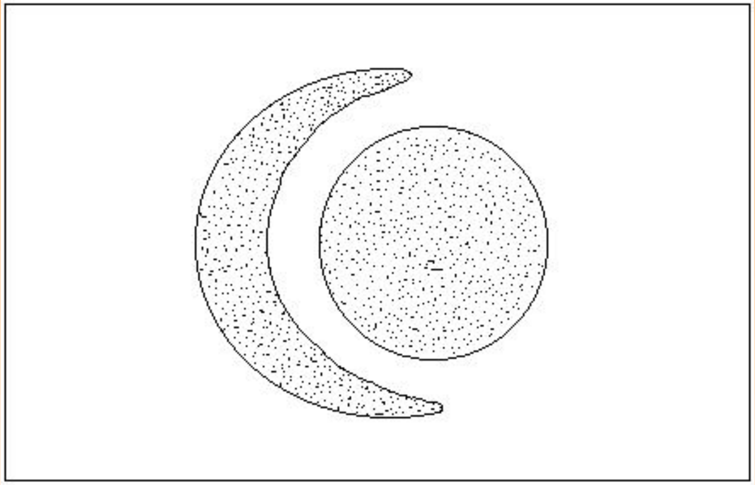
Clustering by the proposed method



Scatter diagram of synthetic data with noise



Scatter diagram with seed points



Clustering by the proposed method

Thank You