PROPERTIES OF POTENTIAL FUNCTION- BASED CLUSTERING ALGORITHMS

Viorel NICOLAU, Gheorghe PUSCASU and Rustem POPA

"Dunarea de Jos" University of Galati
Automatic Control and Electronics Department
47 Domneasca Street, Galati 6200, ROMANIA
Email: vionic@ac.ugal.ro

Abstract: The clustering algorithms based on potential functions are capable of clustering a set of data, making no implicit assumptions on the cluster shapes and without knowing in advance the number of clusters. They are similarity-based type clustering algorithms and do not use any prototype vectors of the clusters. In this paper, some properties of these algorithms are studied: points arrangement tendency, constant potential surface, cluster shapes and robustness to noise.

Keywords: potential function, clustering algorithm, measure of similarity

1. INTRODUCTION

1.1 Clustering algorithms

A data set clustering can be done in two main ways: hierarchical and partitive approaches. The hierarchical methods include agglomerative and divisive algorithms, corresponding to bottom-up and top-down strategies to build a hierarchical clustering tree, which can be used for interpretation of the data structure (Vesanto and Alhoniemi, 2000).

Partitive clustering algorithms divide a data set into a number of clusters according to a generic inter-point measure of similarity or dissimilarity, trying to obtain an optimum value of a performance criterion.

The most commonly used classes of partitive algorithms are similarity-based methods. These methods include algorithms based on distance of a point to the prototype vectors, such as *k-means* and *ISODATA* and potential function-based algorithms.

The algorithms based on potential function use a measure of similarity created with a function between two points of the data set, called potential function, which is a non-increasing function with the distance between the points.

1.2 Optimal clustering and validity indices

In general, optimal clustering means partitioning a data set into a set of clusters, which minimizes distances within and maximizes distances between clusters. However, within- and between- cluster distances can be defined in several ways. In Table 1, within-cluster distances are shown, for a cluster Q_k.

Table 1 Within-cluster distances $d(Q_k)$

Within-cluster distance	$d(Q_k)$
Average distance	$d_a = \dfrac{\sum_{i,j} d(x_i, x_j)}{N_k \cdot (N_k - 1)}$
Nearest neighbor distance	$d_{nn} = \dfrac{\sum_i \min_j (d(x_i, x_j))}{N_k}$
Centroid distance	$d_c = \dfrac{\sum_i d(x_i, c_k)}{N_k}$

N_k represents the number of vectors in cluster Q_k. Also, x_i, $x_j \in Q_k$, $i \neq j$ and c_k is the center of gravity of Q_k:

$$c_k = \sum_{i=1}^{N_k} x_i / N_k \qquad (1)$$

In Table 2, distances between clusters are shown, for two clusters Q_i and Q_k.

This paper was recomended for publication by E. Ceanga

Table 2 Between-cluster distances D(Qi, Qk)

Between-cluster distance	D(Qi, Qk)
Single linkage	$D_m = \min_{i,k}(d(x_i, x_k))$
Complete linkage	$D_M = \max_{i,k}(d(x_i, x_k))$
Average linkage	$D_a = \dfrac{\sum_{i,k} d(x_i, x_k)}{N_i \cdot N_k}$
Centroid linkage	$D_c = d(c_i, c_k)$

N_i and N_k represent the number of vectors in clusters Q_i and Q_k, $x_i \in Q_i$, $x_k \in Q_k$.

To select the best one from many partitions, a validity index can be used to evaluate them. Different validity indices can be defined (Bezdek, 1998), depending on which distances are considered.

For example, the Davies-Bouldin index uses d_C as within-cluster distance and D_C as between-cluster distance. In this case, the best clustering minimizes the expression:

$$\frac{1}{C} \cdot \sum_{i=1}^{C} \max_{i \neq k} \left(\frac{d_c(Q_i) + d_c(Q_k)}{D_c(Q_i, Q_k)} \right) \qquad (2)$$

where C is the number of clusters. This index is suitable for evaluation of partitions with spherical clusters, the best partition being indicated by the index with the minimum value.

2. POTENTIAL FUNCTION-BASED ALGORITHMS

2.1 Potential functions

Consider a data set S of *N* input vectors into a *d*-dimensional space:

$$S = \left\{ x_i \mid x_i = (x_{1i}, x_{2i}, ..., x_{di})^T \in \Re^d, i = \overline{1, N} \right\} \quad (3)$$

A potential function $K(x_i,x_k)$ associated with the vector $x_i \in S$ defines a positive value, called potential of the point x_i to the reference point $x_k \in \Re^d$. The potential depends on distance between the points x_i and x_k, denoted $d_{ik} = d(x_i,x_k)$ and is a non-increasing function with d_{ik} .

Two potential functions are commonly used :

$$K_1(x_i, x_k) = \frac{1}{1 + \alpha \cdot d_{ik}^2} \qquad (4)$$

$$K_2(x_i, x_k) = \exp(-\alpha \cdot d_{ik}^2) \qquad (5)$$

where parameter α controls the slope of the function. The potential values belong to range (0, 1] and the maximum is obtained for $d_{ik} = 0$. The functions are smoothly if parameter α has small values.

The function variations with d_{ik} for different values of α are illustrated in Figure 1. The potential functions K_2 are represented with continuous lines and K_1 are represented with dotted lines.

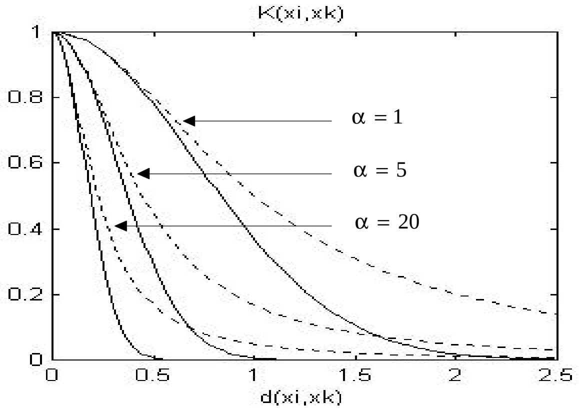

Fig. 1. Potential functions for three values of α

In Figure 1, for the same α value, the two potential functions have similar values if distances between the points are small.

The distance d_{ik} can be the general Minkovski distance:

$$d(x, y) = \sqrt[p]{\sum_{i=1}^{d} |x_i - y_i|^p} \ , \quad x, y \in \Re^d \qquad (6)$$

where for p=2 Euclidean distance is obtained, which is considered in this paper.

A constant potential value to a reference point $x_k \in \Re^d$ is obtained by the potential function $K(x_i,x_k)$ associated with the points $x_i \in \Re^d$ for which the distance d_{ik} is constant. The points x_i generate a constant potential surface, whose shape depends on distance definition. For Euclidean distance, the constant potential surface has spherical shape around the reference point $x_k \in \Re^d$.

The parameter α also affects the constant potential surface, different α values generating different potential surfaces, but their shapes are similar around the reference point. If α value increases, the potential surface is moving nearer to the reference point.

Similar, a potential value of a point x_i to a group of reference points $M=\{x_{k1}, x_{k2},..., x_{km}\}$ can be defined as the average of the potential values of the point x_i to all reference points x_{kj}:

$$A_i = A(x_i, M) = \frac{1}{m} \cdot \sum_{j=1}^{m} K(x_i, x_{kj}) \qquad (7)$$

In this case, a constant potential value to the group M generates a potential surface, which also depends on distance definition. The constant potential surfaces surround the reference points, but their shapes are affected by α-values and reference point positions.

For example, two reference points x_{k1}, $x_{k2} \in \Re^2$ and a constant potential value K = 0.5 are considered. The reference points are represented with '+' in Figure 2.

For every reference point, three different constant potential surfaces are generated with potential function $K_2(x_i,x_k)$, corresponding to K and three α-values: 2, 5 and 10.

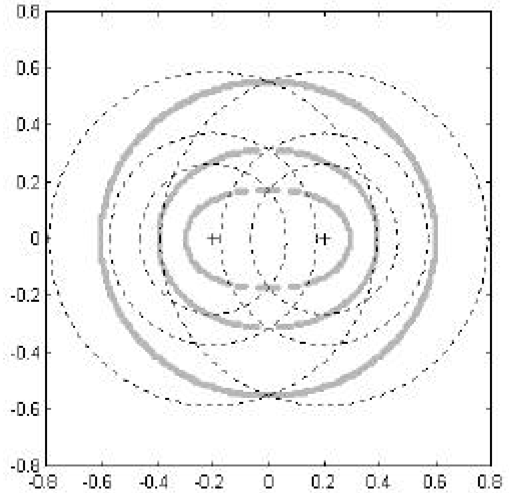

Fig. 2. Constant potential surfaces in $\Re^2$ space

In Figure 2, the constant potential surfaces to a single reference point have spherical shapes around the point, being represented with dotted lines. The big circle represent the first potential surface, corresponding to α=2.

The potential surfaces generated by the constant potential value K to the reference group $\{x_{k1}, x_{k2}\}$ are represented with continuous lines in Figure 2, corresponding to the same α-values. They surround the group of reference points, but the shapes depend on α-values and reference point positions. For small α-values, the shape tends to be spherical.

2.2 The algorithm stages

A potential function-based algorithm (*PFBA*) uses a measure of similarity, which characterizes the membership of a point to a group of points, based on a potential function (Dorofeyuk, 1966).

Consider a group of points M from S, $M \subset S$ and a point $x_i \in S$, $x_i \notin M$. A similarity measure of x_i to M can be defined as the average A_i of the potential values of the point x_i to all points of the group M:

$$A_i = A(x_i, M) = \frac{1}{N_M}\cdot\sum_{x_j\in M} K(x_i, x_j) \qquad (8)$$

where N_M represents the number of points in M.

Using this measure of similarity, the points of the data set S can be arranged in a certain order, starting from a specified point, pursuant to the following rule:

- select the starting point, let it be $x^1 \in S$, form the first group $M_1=\{x^1\}$ and denote $A_1=1$, which represents the maximum potential value;
- find in S/M_1 the point x^2 with the maximum measure of similarity to M_1 in the meaning of (8), which is:

$$A_2 = A\left(x^2, M_1\right) = \max_{x\in S/M_1}\left(A(x, M_1)\right) \qquad (9)$$

Form a new group $M_2 = \{M_1, x^2\} = \{x^1, x^2\}$;

- repeat the previous step until all the points of the data set S are assigned, by finding the points x^k with maximum measure of similarity to M_{k-1}:

$$A_k = A\left(x^k, M_{k-1}\right) = \max_{x\in S/M_{k-1}}\left(A(x, M_{k-1})\right) \qquad (10)$$

Form the groups $M_k=\{M_{k-1}, x^k\}$.

In this way, the set S is ordered, $S=\{x^1, x^2,..., x^N\}$ and a new series is obtained: $A_1, A_2,...., A_N$.

All potential function-based algorithms compute the new series $A_1,..., A_N$, which contains the necessary information for clustering. The analysis of this series differs from algorithm to algorithm.

For example, the algorithm considered in this paper (Bumbaru, 1970) has the following stages:
- select the starting point;
- arrange the points of the data set S, using the rule described above;
- compute the ratios $R_1,..., R_N$, where

$$R_1 = 1, \quad R_k = \frac{A_{k-1}}{A_k}, \quad k = \overline{2, N} \qquad (11)$$

- compute the mean value m_R and the standard deviation σ_R of the ratios R_k;
- consider a threshold $p = r\cdot c\cdot\sigma_R$, where r = 1...20 and $c\in[0.3, 1]$;
- the clustering decision is made comparing the difference R_k-R_{k-1} with the threshold and a new cluster begin if $R_k - R_{k-1} > p$. Thus, a new partition is obtained;
- compute other partitions for different threshold values, by increasing *r*, until $p > R_k - R_{k-1}$ for all differences.

The clustering result is considered the partition, which remains unchanged for the greatest number of *r*-values.

3. PROPERTIES OF *PFBA*

3.1 Points arrangement tendency

Arrangement of the points in the ordered data set S depends on selections of first point, potential function and parameter α and the tendency is to order the points in successive layers around the first points. Also, these selections affect the values of series A_k and R_k and can affect the clustering performance.

To illustrate the influence of the first point selection and the ordering tendency of the points, a data set is clustered starting from two different points.

Consider the data set I, with two spherical and well-separated clusters, as shown in Figure 3. The clusters have 40 and respectively 100 points, which are denoted x_1, ...x_{140}, marked with '+' in the figure.

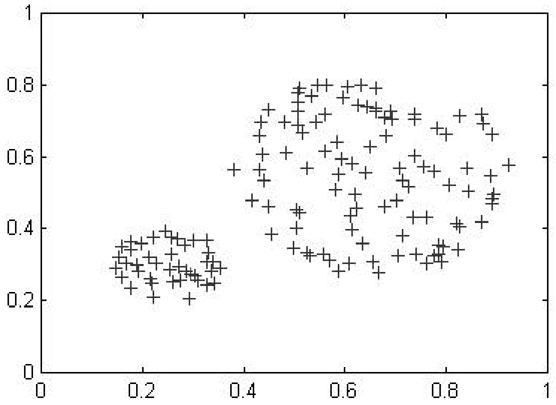

Fig. 3. Data set I, with two spherical clusters

Running the clustering algorithm twice, with the first point into the small cluster and then into the big cluster, the arrangements of the data set points are different and are illustrated in Figures 4 and 5. The potential function is K_1 and the parameter α=10.

In these figures, the points are marked with '+' and the start point is marked distinctly. In addition, lines are drawn between every two consecutive points in the ordered data set.

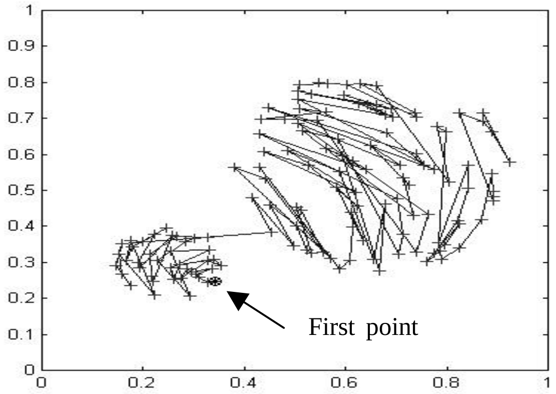

Fig. 4. The ordered data set, starting from x_1

In Figure 4, the arrangement starts with the first point of the data set x_1, which is into the small cluster and in Figure 5 the start point is x_{120}, which is into the big cluster.

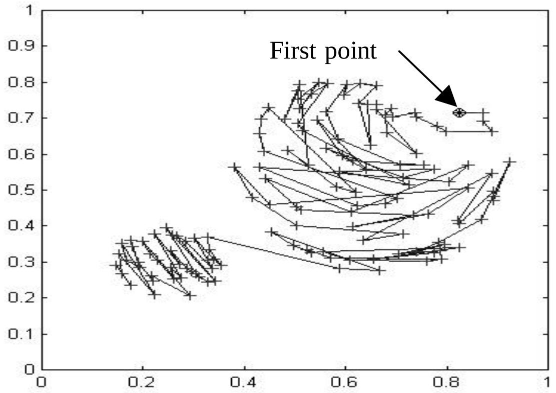

Fig. 5. The ordered data set, starting from x_{120}

It can be observed that the points are ordered in successive layers around the first ordered points.

3.2 The series A_k and R_k

The series A_k has decreasing tendency, representing the average of the potential values of a point x_i to all points placed before it in the ordered data set. Big variations between adjacent elements of A_k indicate the transition into another cluster.

The ratio R_k can be considered a random variable, with standard deviation σ_R and mean value m_R close to 1 for any data set containing sufficiently great number of points and for a large range of parameter α. The transition from one cluster into another is indicated by big values of R_k.

The series A_k and R_k computed for the situations above are illustrated in Figures 6 and 7. The first row represents the series A_k and the second row represents the series R_k.

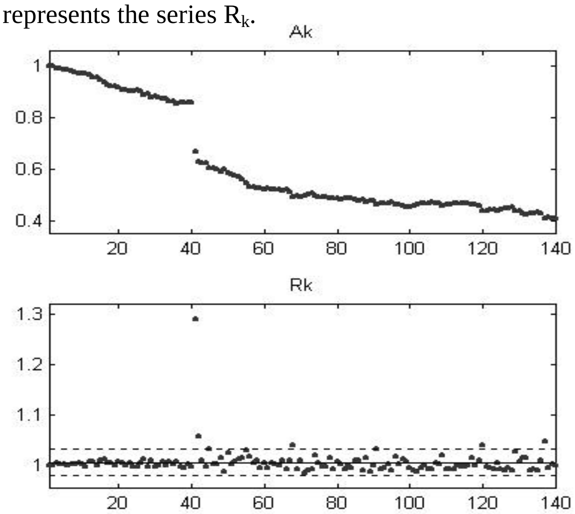

Fig. 6. The series A_k and R_k with the start point x_1

In R_k window, the mean value m_R is represented with continuous line and the $m_R \pm \sigma_R$ values are illustrated with dotted lines.

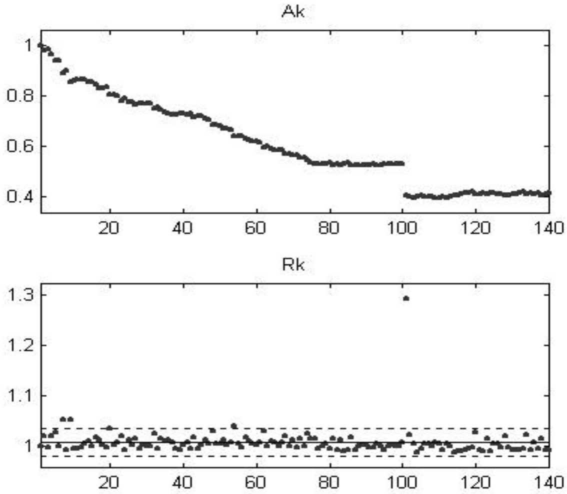

Fig. 7. The series A_k and R_k with the start point x_{120}

The series values are different and the maximum variations of them indicate the cluster separation, which is evident. However, the first separation is bigger and the arrangement is more appropriate.

3.3 Constant potential surface

The constant potential surface to a group of points M depends on potential function and parameter α and tends to take similar shape as the one of the cluster when α increases, even for more complex cluster.

The influence of α values on constant potential surface is illustrated for two different values of α, using the potential function K_2. Increasing α value, the constant potential surface will be closely to the cluster points and the new cluster will be oblong. Thus, the parameter α can be used to characterize the shape of the clusters: more compact or oblong.

Consider a complex cluster M with 199 points and a new point x_{200}, which has the measure of similarity to M denoted A_{200}. The value of the constant potential surface was chosen equal to A_{200}, which is useful to compare new additional points with x_{200}.

For α=25, the constant potential value is A_{200}=0.055 and the constant potential surface is illustrated with gray color in Figure 8. The points of the cluster are marked with '+' and the last point placed on the constant potential surface is marked with 'o'.

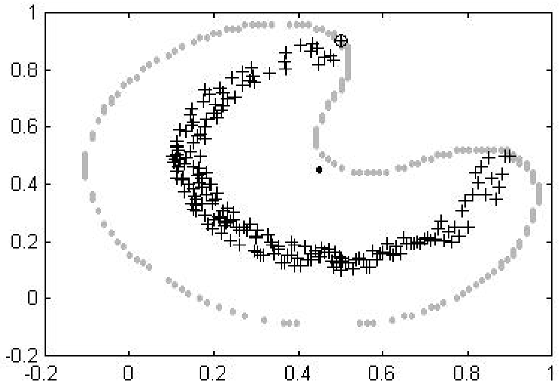

Fig. 8. Constant potential surface for α = 25

Similar, for α = 80, the constant potential value is A_{200} = 0.029 and the constant potential surface is illustrated with gray color in Figure 9.

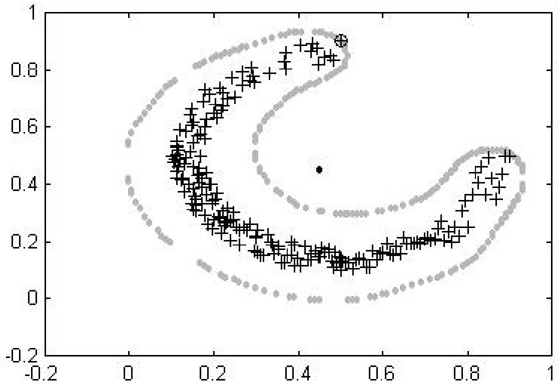

Fig. 9. Constant potential surface for α = 80

Additional points placed outer potential surface have measure of similarity to M smaller than A_{200} and the points are ordered after x_{200}.

By contrary, any additional point placed into potential surface is ordered before x_{200}.

For example, in Figures 8 and 9, the point at the (0.45, 0.45) coordinates is marked with '.'. This point is ordered before x_{200} if α = 25 and is ordered after x_{200} if α = 80.

3.4 Cluster shapes

The potential function-based algorithms work well for complex cluster shapes. In contrast, the algorithms based on distance to the prototype vectors are sensitive to the cluster shapes and give good results just for spherical well-separated clusters.

Two cases are considered: elongated and irregular shapes of the clusters. In the first case, the data set II with two elongated clusters is chosen. The clusters have 50 and respectively 100 points and their main directions are parallel. Using *PFBA*, the clusters are well identified, as shown in Figure 10.

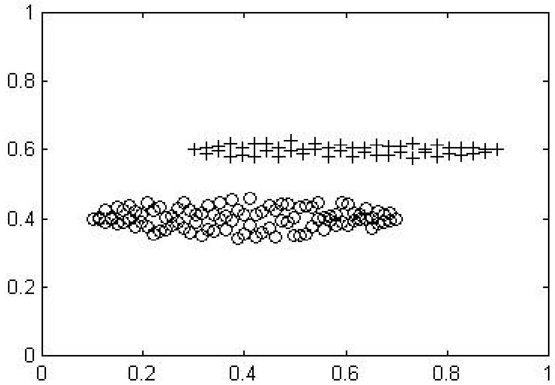

Fig. 10. Clustering elongated shapes with *PFBA*

The potential function K_2 was used, with parameter α = 60. The ordered data set is obtained starting from x_{60}, which is marked distinctly, as illustrated in Figure 11.

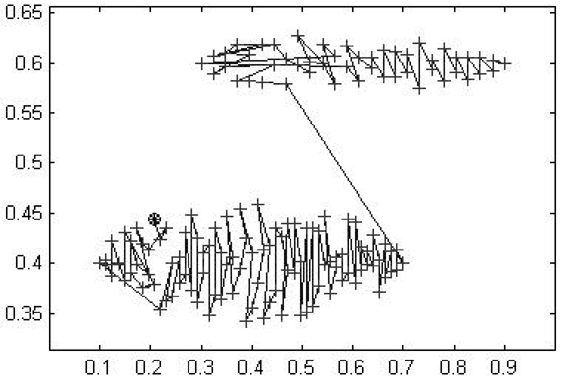

Fig. 11. The ordered data set II, starting from x_{60}

The boundary between clusters can be easily detected by analyzing the series A_k and R_k, which are represented in the Figure 12. In the second series, it is easier to identify the clusters, because the cluster separation in R_k is bigger.

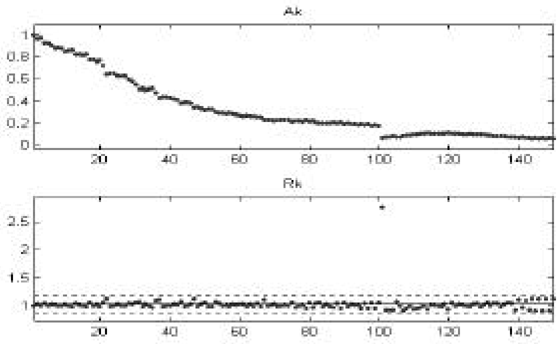

Fig. 12. The series A_k and R_k for ordered data set II

In the second case, the data set III with two irregular clusters is considered. The clusters have also 50 and respectively 100 points. Even for different starting point, the clusters are well identified with *PFBA*, as shown in Figure 13.

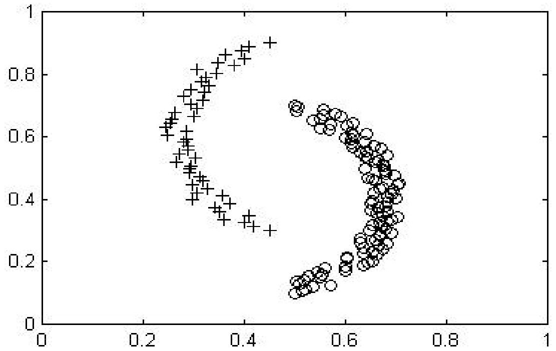

Fig. 13. Clustering irregular shapes with *PFBA*

The parameter α must characterize oblong clusters and its value ought to be big, being chosen α = 60. Two starting points were chosen, x_1 and x_{60}, and the ordered data sets are illustrated in Figures 14 and 15.

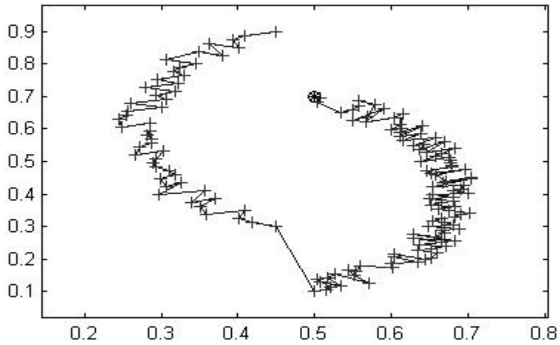

Fig. 14. The ordered data set III, starting from x_1

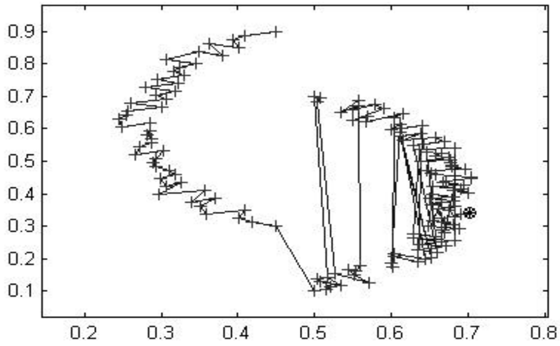

Fig. 15. The ordered data set III, starting from x_{60}

3.5 Robustness to noise

In many cases, data sets are affected by noise, which can radically change the clustering results, by modifying the positions of set points.

Consider the data set IV, with two small clusters, which are well identified by both clustering algorithms: *PFBA* and *ISODATA*. In Figure 16 are illustrated: the clusters, the cluster centroids and the boundary between them.

If the noise affects the data set and changes the position of one point which is marked with '+', the *PFBA* clustering is not affected, but the *ISODATA* clustering is modified, as illustrated in Figure 17.

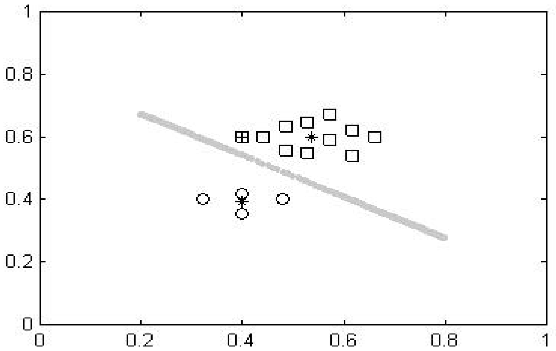

Fig. 16. Clustering of the data set IV, with *PFBA*

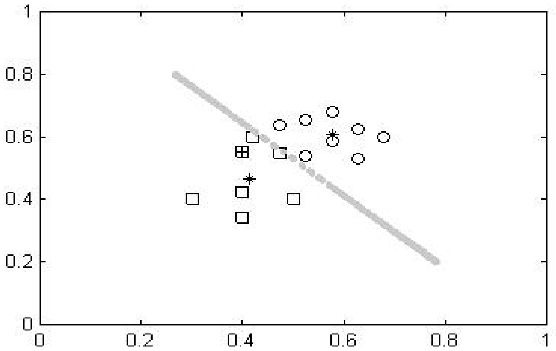

Fig. 17. Noisy set clustering with *ISODATA*

4. CONCLUSION

The *PFBA* do not use any prototype vectors of the clusters. Therefore, they give good results even for complex shape clusters. In addition, *PFBA* can separate singular points and are more robust to noise.

REFERENCES

Bezdek, J.C. (1998). Some new indexes of cluster validity. *IEEE Trans. Syst., Cybern.*, **28**, 301-315.

Bumbaru, S., E. Ceanga and I. Bivol (1970). Self-learning classifier for data analysis. In: *VI-th Iugoslav Int. Symp. on Inf. Processing*, **H5**, 1-6.

Dorofeyuk, A.A. (1966). Algorithms of teaching the machine the pattern recognition without teacher based on the method of potential functions. *Avtomatika i Telemechanica*, **10**, 78-87.

Vesanto, J. and E. Alhoniemi (2000). Clustering of the Self-Organizing Map. *IEEE Trans. on Neural Networks*, **11**, 586-600.