Method of fuzzy clustering with genetic algorithm

Novoselova N.A.[1), Tom I.E.[1)

1) United Institute of Informatics Problems, NAS Belarus, Surganova str. 6, 220012 Minsk, Belarus
e-mail: {novosel, tom}@newman.bas-net.by

Abstract: *the original method of fuzzy clustering using genetic algorithm is proposed. The chromosomes of the population of genetic algorithm have variable length and encode the possible partition of the data. The method enables in one GA run to find the near optimal data partition into clusters, to determine the cluster number and simultaneously to speed up clustering process comparing to other known methods. Classification rules, constructed on the basis of clustering results, obtained by proposed method, provide more accurate classification in comparison with results of classical FCM-method.*

Keywords**:** Clustering, classification rule, genetic algorithm.

1. INTRODUCTION

The clustering task consists in data partitioning into several subsets (clusters) of objects, which are similar in some sense, taking into account its description in feature space. In metric space the "similarity" is usually defined by distance. The distance can be calculated between initial objects as well as between these objects and cluster centers. Usually the coordinates of prototypes are unknown in advance – they are determined together with data partition into clusters..

There exists a great amount of clustering methods, which can be divided into the exact clustering [1-Айвазян, 2-Мандель] and fuzzy clustering approaches [3-Бэждек,4-Бэждек, 5-Gustafson, 6-Gath, 7-Hoppner]. The methods of exact clustering partition the initial set of data objects into several non-overlapping subsets. In that case any data object belongs only to single cluster. Methods of fuzzy clustering enable the same object belonging simultaneously to several (or even to all) clusters, but with different membership degree. In many cases the fuzzy clustering is more "natural" than the exact clustering, for example for the object at the clusters' boundaries. Cluster methods are widely used for the analysis of multivariate data in medical, industrial, financial and other applications. In these applications it is necessary to reveal the groups of objects with similar features and behavior and on the basis of defined groups to make a decision about new data object by estimating its similarity to one or another previously defined group. One of the major problems in clustering is the determination of the number of clusters in data. That cluster number specifies the further course of clustering algorithm and as a rule is unknown in advance. Therefore the development of methods for automatic determination of the cluster number during the clustering process is of particular importance. In the following sections one of the approaches to the solving of that problem is considered.

2. QUALITY MEASURES FOR ESTIMATION OF FUZZY CLUSTERING RESULTS

The fuzzy partition $P = \left\{ A^1, ..., A^c \right\}$ of data objects from data set $X = \left\{ x_1, ..x_n \right\}$ into predefined number of fuzzy clusters *c* represents the solution of fuzzy clustering task. Fuzzy clusters are described by the following fuzzy partition matrix:

$$U = \left[\mu_{ij} \right], \quad \mu_{ij} \in [0,1], \quad i = \overline{1,c};\ j = \overline{1,n}\,, \qquad (1)$$

where the *i*-th column constitutes the membership degrees of object $\mathrm{x}_i = \left(x_i^1, \ldots, x_i^m \right)$, defined by vector of feature values, to the clusters $A^1, ..., A^c$. Moreover the following conditions must be satisfied:

$$\sum_{i=1}^{c} \mu_{ij} = 1,\ j = \overline{1,n}; \quad 0 < \sum_{j=1}^{n} \mu_{ij} < n\,, \qquad (2)$$

The fuzzy clustering task consists in finding the optimal value of some functional $F(P)$ on the set of all fuzzy partitions of data objects:

$$F(P) \rightarrow \underset{P \in U}{extr}\,, \qquad (3)$$

where *U* – the set of all possible fuzzy partitions.

The more frequently used measures, characterizing the quality of data partition into groups using fuzzy clustering algorithms are described in [7].

Most of researchers consider the quality measure of data partition as a mapping *Q* of the set of partition matrices *U* into the real number range - $Q : U \rightarrow R$. Extreme value of function *Q* defines the optimal number of clusters or data groups.

In order to find the optimal value of the quality measure of data partition the fuzzy clustering algorithm must be performed several times, every time with different predefined number of clusters. Then the local optimum of quality measure helps defining the more appropriate number of clusters in the analyzed data.

The most frequently used quality measures of data partition are the partition index, partition entropy and Xie-Beni index [8-Хи-Бени].

Partition index. Partition index for matrix *U*, denoted as *P*(*U*), is defined as mean value of square membership function values, appeared in fuzzy partition matrix:

$$P(U) = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^2 \qquad (4)$$

When every object belongs only to single cluster (in the case of exact clustering), then the partition index has a maximal value and is equal to 1. If the object equally belongs to all clusters and membership value to every cluster is 1/c , then the partition index $P(U)$ has minimal value and is equal to 1/c.

Partition entropy. The partition entropy value is defined in the following way:

$$H(U) = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}\ln(\mu_{ij}) \ . \qquad (5)$$

Partition entropy values lie in the range $\left[0, \ln(c)\right]$, while the maximal value corresponds to uniformly distributed membership values of data object to all clusters (=1/*c*):

$$H(U) = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j}^{n}\frac{1}{c}\ln(\frac{1}{c}) = \ln(c) \ .$$

Though the above described quality measures of data partition are applicable for the estimation of partition matrix, their use for the determination of number of clusters is not apparent. To the drawbacks of these quality measures the following can be ascribed:

1) The monotone dependence on the number of clusters;

2) Sensitivity to the value of the fuzziness index γ. When $\gamma \to 1$, the quality measures are equal for all number of clusters *c*, when $\gamma \to \infty$ both above described quality measures define as an optimal the data partition into two clusters;

3) direct relation to the data geometry, i.e. to the data distribution in multivariate feature space is absent. The data themselves are not used in the calculation of values of quality measures.

As the quality measures of data partition, which use both the membership values and data values, the following measures can be considered: Xie-Beni index, Fukuyama-Sugeno index, quality indices in [7-Hoppner, 9-Смит]. In this paper only the Xie-Beni index will be considered.

Xie-Beni cluster validity index is defined as an estimation function of cluster's compactness and separation. For fuzzy partition $U=[\mu_{ij}]$ fuzzy deviation of the data object x_j from cluster *i* is defined in the following way:

$$d_{ij} = \mu_{ij}\left\|x_j - \tau_i\right\|$$

Variation σ_i for cluster *i* is defined as a sum of squared fuzzy deviations of the data objects from *X*. The total variation of clusters σ is equal to the sum of variations σ_i. for all clusters. The value $\pi = \left(\sigma_i / n_i\right)$ denotes the compactness of cluster *i*. As n_i equals the number of objects, belonging to cluster *i*, then π - the mean variation in cluster *i*.

The separation value of fuzzy clusters is defined as minimum of distances between cluster centers:

$$d_{\min} = \min\left\|\tau_i - \tau_j\right\|$$

Thus the Xie-Beni index is defined in the following way:

$$XB = \pi / n \cdot d_{\min} \ ,$$

where *n* – the number of data objects.

The small values of Xie-Beni index corresponds to compact and well separated clusters. However the value of Xie-Beni index monotone decrease with the increase of the number of clusters. In order to exclude the impact of this decreasing tendency of the index on the estimation of optimal number of clusters, the limit of cluster number $c_{\max}$ is determined for the analyzed data and the search of the minimal value of Xie-Beni index is performed in the range $\left[1, c_{\max}\right]$ by the repeated running of clustering algorithm.

All the above considered and only mentioned quality measures of data partition have the common essential deficiency – they require the repeated solving of clustering task and therefore are inefficient for many practical applications.

For the purpose of solving the task of determination of the cluster number in one run of clustering algorithm the authors have developed the method of fuzzy clustering, which uses the search capabilities of genetic algorithm (GA). Genetic algorithm enables in the process of population evolution the solving of clustering optimization task simultaneously with determination the near optimal number of clusters in data. [10]. For this purpose the quality measure of data partition - Xie-Beni index is applied.

3. DESCRIPTION OF THE METHOD OF FUZZY CLUSTERING WITH GENETIC ALGORITHM

As the task of fuzzy clustering can be considered as the nonlinear optimization task, the authors have attracted the genetic algorithm to find the clusters in data. The proposed method of fuzzy clustering on the basis of GA uses the chromosomes with variable length and enables searching for the clustering result, more efficient according to Xie-Beni cluster validity index. Each chromosome encodes a possible partitioning of the data by means of cluster centers. The advantage of proposed approach to fuzzy clustering is the possibility in one run of GA to find not only the optimal data partitions into clusters, but also to determine their number, which corresponds to the minimal value of cluster validity index. Denote the set of all possible fuzzy partitions as *U*, where

$$U = \left\{F \in R^{c\times n} \mid \sum\nolimits_{i=1}^{c}\mu_{ij} = 1,\ 0 < \sum\nolimits_{j=1}^{n}\mu_{ij} < n,\ \mu_{ij} \in \left[0,1\right]\right\},$$

The best matrix of fuzzy partition F^* is defined by the following:

$$F^* \in U, \quad XB\left(F^*, V^*, X\right) = \min_{F^* \in U} XB\left(F, V, X\right),$$

where *XB* – the value of Xie-Beni index, *V* – the set of cluster centers, *X* – the set of data objects.

The number of clusters together with the corresponding cluster coordinates evolve simultaneously in the process of genetic algorithm execution. As the chromosome length is variable, a single population can contain the individuals, encoding the different number of cluster centers.

The particular individuals of population represent the solution of clustering task and they are made up of real numbers which define the coordinates of the cluster centers $\tau_i \in V$, i=1,…,c of partition. The length of i-th individual equals $L_i = c \cdot m$, where c – the number of clusters, m – the dimensionality of data objects or the number of features. For example, for 3-dimensional data objects, the chromosome (1.1, 4, 6.3)(7.5, 2.4, 3.8) presents the coordinates of two cluster centers. Each individual of GA population initially encodes the centers of $c \in [2, max]$ clusters, where *max* – predefined maximal number of clusters in the chromosome of initial population. During evolution the length of chromosome can vary arbitrarily, and later on is not limited by the value *max*. The initial population is generated at random using the data objects. The cluster centers τ_i, which are encoded in chromosomes, present the coordinates of data objects, which are at random selected from data set.

The value of fitness function indicates the degree of goodness of the GA individual and the efficiency of the solution of optimization task. The inverse value to the Xie-Beni index is used as the fitness function in our GA. The Xie-Beni index is defined as a function of the ratio of the total variance σ of the data objects to the minimum separation *div* of the clusters:

$$\sigma(U,V,X) = \sum\nolimits_{i=1}^{c} \sum\nolimits_{j=1}^{n} \mu_{ij}^2 d(\mathrm{x}_j, \tau_i{}^k)$$

$$div(V) = \min_{i \neq j} \{ \| \tau_i - \tau_j \|^2 \},$$

$$XB(U,V,X) = \frac{\sigma(U,V,X)}{n \cdot div(V)},,$$

where c – the number of clusters, n – the number of data objects, $d(\mathrm{x}_j, \tau_i) = \| \mathrm{x}_j - \tau_i \|^2$. The fitness function is calculated as $f = 1/XB(U,V,X)$. Optimal fuzzy partition corresponds to the maximal value of function f, moreover the value σ is sufficiently small, whereas the separation value of cluster centers *div* – sufficiently high.

For the generation of the population of individuals of GA the following steps are performed:

1. Initially the coordinates of cluster centers are encoded in the chromosomes. Then the values of memberships μ_{ij} of each data object to each cluster are calculated:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d(\mathrm{x}_j, \tau_i)}{d(\mathrm{x}_j, \tau_k)} \right)^{\frac{1}{\gamma - 1}}} \quad 1 \le i \le c, \;\; 1 \le j \le n$$

When even so the value $d(\mathrm{x}_j, \tau_i)$ for some cluster center τ_i equals zero, then μ_{ij} =1 (k=i) and μ_{ij} =0 for k=1,…,c and k≠i.

2. The values of cluster centers are updated according to the following expression:

$$\tau_k = \frac{\sum_{j=1}^{n} (\mu_{kj})^{\gamma} \mathrm{x}_j}{\sum_{j=1}^{n} (\mu_{kj})^{\gamma}}, \quad 1 \le k \le c.$$

3. For each individual of GA the fitness function value f is calculated.

4. For the GA operation of selection of the individuals the ordinal proportional sampling is applied. The number of copies of each individual in the new population is proportional to the value of its fitness function, i.e. more adapted individuals survive and are copied into the new population..

5. For the modification of the particular individuals of the population the recombination operations are performed.

During the execution of GA crossover operation of parent individuals the cluster centers are considered indivisible, i.e. the crossover point can locate only between the coordinates of two adjacent clusters. Moreover during crossover operation the child individuals are ensured to encode at least two cluster centers. Thus, the crossover operation is executed in the following way:

1) suppose, that two parent individuals i_1 and i_2 consists of M_1 and M_2 clusters accordingly;

2) define the crossover point for the parent individual i_1 as $\tau_1 = rand() \bmod M_1$, i.e. τ_1 can take the values from 0 to M_1-1. For the possibility the child individuals encode at least two cluster centers, the crossover point τ_2 for the second individual i_2 must be located in some range [L_{bound}, U_{bound}], where L_{bound} and U_{bound} are calculated according to expression:

$$\begin{aligned} L_{bound} &= \min\left(2, \max\left(0, 2 - \left(M_1 - \tau_1\right)\right)\right) \\ U_{bound} &= \left(M_2 - \max\left(\left(0, 2 - \tau_1\right)\right)\right) \end{aligned};$$

3) further the crossover point for the second individual τ_2 is determined in following way:

$$\tau_2 = \begin{cases} L_{bound} + rand() \bmod \left(U_{bound} - L_{bound} + 1\right), \\ 0, \end{cases}$$

$$\text{если} \quad L_{bound} \le U_{bound}$$
$$\text{если} \quad L_{bound} > U_{bound}$$

After execution of crossover operation with probability P$_{cross}$, the mutation operation of separate genes (cluster coordinates) of new crossed individuals is performed with probability P_{mute}. At first the random value τ is generated, which is uniformly distributed in the range $\tau \in [0,1]$. During mutation operation the coordinate value, which is encoded in single gene is updated in such a way:

$$v^* = \begin{cases} (1 \pm 2 \cdot \tau) \cdot v, & \text{если} \quad v \neq 0 \\ \pm 2 \cdot \tau, & \text{если} \quad v = 0 \end{cases}$$

The arithmetic operations '+' or '–' are applied equiprobable. During the execution of mutation operation the coordinate values can be modified for more than one cluster center.

6. The best individual of previous population is copied unchanged into the new population, which corresponds to the GA elitism operation.

All the above listed steps are executed the finite number of times until the GA stopping condition will be satisfied. As a stopping condition either the pre-defined, maximal number of generations, or the unchangeable cluster coordinates in the best, according to fitness value, individual can be used.

The results of fuzzy clustering have been used for the construction of the set of fuzzy classification rules. The membership functions of antecedents and consequence of each rule are defined using the obtained fuzzy partition of data objects in the *p*-dimensional feature space. Each obtained cluster determines a single fuzzy rule. From the mathematical point of view the membership degree of the feature value *y* to the *p*-th projection $\mu_k^{(p)}$ of the fuzzy cluster *k* equals the supremum of all the membership degrees of data objects to cluster *k*, *p*-th component of which is equal to *y*:

$$\mu_k^{(p)}(y) =$$
$$= \sup\left\{ \frac{1}{\sum_{j=1}^{c}\left(\frac{d^2(\tau_k, \mathrm{x})}{d^2(\tau_j, \mathrm{x})}\right)^{\frac{1}{\gamma-1}}} \middle| \mathrm{x} = \left(x_1, \ldots, x_{p-1}, y, x_{p+1}, \ldots, x_m\right) \in R^m \right\} \quad (6)$$

As the calculation of all membership degrees according to expression (6) is sufficiently complex, then the more simple procedure is usually applied. According to that procedure the membership functions of the rule antecedent are generated by the pointwise projection of the fuzzy partition matrix onto the one-dimensional coordinate features space [11], which results in the one-dimensional discrete fuzzy sets. For the transformation of discrete membership function to the continuous function its convex envelope is calculated. The envelope is further approximated by triangular or trapezoid function using the heuristic algorithm, which minimizes the error sum of squares [12].

4. VERIFICATION OF THE EFFICIENCY OF CLASSIFICATION RULES, CONSTRUCTED ON THE BASIS OF PROPOSED CLUSTERING METHOD

Fuzzy rules are introduced in the following way:

R_i: When $x \in A$, then class C_1 with weight p_{i1} and … and class C_M with weight p_{iM} , where *i*=1,…,*c*, *A* – fuzzy set, defined by multivariate membership function, corresponding to fuzzy cluster; $x = (x^1, ..., x^p)$ - *p*-dimensional data object; *M* – the number of class labels; C_j – class label in the consequence of the rule (*j*=1,…,*M*); p_{ij} – the confidence degree of the rule R_i.for class C_j .For the determination of class label for the new data object using the set of fuzzy rules the following reasoning mechanism is applied:

Algorithm 1

1) the membership degree μ_{ik} of data object x_k to each fuzzy cluster *i* ($i = 1, \ldots, c$) is calculated using the following expression:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{\gamma-1}}} \quad (7)$$

2) for the definition for the data object x_k of the "voice" for each class or the degree of assignment of the data object to each from the existing clusters the following expression is used [7]:

$$V_{C_j}(x_k) = \sum_{R_i,\, i=1\ldots,c} \mu_{ik} \cdot p_{ij}, \quad j = 1\ldots, M \quad (8)$$

3) the class label for the data object x_k is defined according to the class label, which has the majority of "voices": $x_k \to C_{j^*}$, where

$$V_{C_{j^*}} = \max\,\{V_{C_j}(x_k) \mid j = 1, \ldots, M\} \quad (9)$$

The confidence degree p_{ij} for each class C_j (*j*=1,…,*M*) and for each rule R_i ($i = 1, \ldots, c$) is defined on the basis of fuzzy clustering results, applied to the data set $X = \{x_1, x_2, \ldots, x_n\}$ and using the following algorithm:

Algorithm 2

1) membership degrees μ_{ik} of all data objects x_k, *k*=1,…,*n* to each cluster *i* ($i = 1, \ldots, c$) are calculated using the expression (7);

2) for each class C_j ($i = 1, \ldots, M$) the sum β_j of membership degrees of data objects to fuzzy rule R_i ($i = 1, \ldots, c$) are calculated:

$$\beta_j(R_i) = \sum_{x_k \in C_j} \mu_{ik}, \quad j = 1, \ldots, M \quad (10)$$

3) confidence degree p_{ij} for each class C_j (*j*=1,…,*M*) and for each rule R_i ($i = 1, \ldots, c$) is calculated with the following expression:

$$p_{ij} = \frac{\beta_j(R_i)}{\sum_{k=1}^{M}\beta_k(R_i)}, \quad i = 1, \ldots, c; \quad j = 1, \ldots, M \quad (11)$$

Proposed method of fuzzy clustering using GA and subsequent procedure of the construction of the set of fuzzy classification rules are tested on the artificially generated data set **set_all** and on the data set **Iris** from the international data archive on machine learning.

During classification of data objects from artificial data set using the obtained fuzzy rules the classification accuracy equals 99.7%, i.e. only one data object is incorrectly classified. In the case of the *FCM*-algorithm for data clustering and further construction of fuzzy rules' set 4 data objects are incorrectly classified. The results of clustering and classification of data set set_all using the proposed method are presented on figure 1.

The second data set Iris consists of three classes (Setosa, Versicolour, Virginica), each class of data includes 50 objects, which are characterized by the values of four features x=(x_1,…,x_4). The class Setosa is linearly separated from two other classes, while the classes Versicolour and Virginica aren't linearly separable. The best result of standard clustering method *FCM*, according to Xie-Beni index, is the clustering result with two clusters. *FCM* clustering algorithm was applied 9 times for each number of clusters in the range $c = \overline{2,\ldots,10}$. After that the fuzzy classification rules have been constructed, and classification of Iris data objects have produced the classification error equal to 10,7%.

Using the proposed clustering method and as a result of 8 from 10 runs of genetic algorithm the data objects were partitioned into 3 clusters with the value of Xie-Beni index equals to 0,15. The classification error was equal to 6,6%, or 4,1% better, than the results, obtained by *FCM* algorithm.

Thus the proposed clustering method enables not only to automate the process of determination the number of clusters, corresponding to data, but also improves the classification accuracy in comparison with application of standard clustering algorithm *FCM*.

The results of clustering and classification for the Iris data set can be seen on figure 2.

4. CONCLUSION

The method of fuzzy clustering with genetic algorithm is proposed. The genetic algorithm encodes the solution of clustering task, using the variable-length chromosomes, which allow finding the near optimal partition of data objects in one run of GA, defining the number of clusters and simultaneously speeding up the clustering process in comparison with known methods. The testing results of proposed method on artificial data set **set_all** и real data set **Iris** from the international data archive of machine learning have indicated the possibility of the new fuzzy clustering method not only to define automatically the number of clusters and to perform the data clustering with the better value of Xie-Beni index in comparison with clustering method *FCM*, but also to construct the fuzzy classification rules, that more accurately classify the data objects.

5. REFERENCES

[1] С.А. Айвазян., В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.

[2] И.Д Мандель. Кластерный анализ. — М.: Финансы и статистика, 1988.

[3] J.C. Bezdek Statistical parameters of cluster validity functionals / J.C. Bezdek, M.P. Windham, R. Elrich // International Journal of Parallel Programming. – 1980. – Vol. 9, № 4. – P. 323–336.

[4] J.C. Bezdek Pattern recognition with fuzzy objective function algorithms / J.C. Bezdek. – New York: Kluwer Academic / Plenum Publishers, 1981. – 272 p.

[5] D.E. Gustafson Fuzzy clustering with a fuzzy covariance matrix / D.E. Gustafson, W.C. Kessel // Proc. of IEEE Conference Decision Control San Diego, CA, 1979. – P. 761–766.

[6] I. Gath Unsupervised optimal fuzzy clustering / I. Gath, A.B. Geva // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1989. – Vol. 11, № 7. – P. 773–781

[7] F. Hoppner et al.. Fuzzy Cluster Analysis– John Wiley and Sons, 1999. – 289 p.

[8] Xie, X.L., Beni, G. A Validity Measure for Fuzzy Clustering // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1991. – Vol. 13, № 4. – P. 841–846.

[9] Smyth, P. Clustering using Monte Carlo Cross-Validation // Proceedings of KDD Conference. –1996.

[10] Novoselova, N. Supervised fuzzy clustering using genetic algorithm for the fuzzy classifier construction /N. Novoselova // Искусственный интеллект. – 2007. – № 4. – С. 343–351.

[11] Klawonn, F. Constructing a fuzzy controller from data / F. Klawonn, R. Kruse // Fuzzy Sets and Systems. – 1997. – Vol. 85, № 2. – P. 177–193.

[12] Sugeno, M. A fuzzy-logic-based approach to qualitative modeling / M. Sugeno, T.Yasukawa // IEEE Transactions on Fuzzy Systems. – 1993. – Vol. 1, № 1. – P. 7–31.

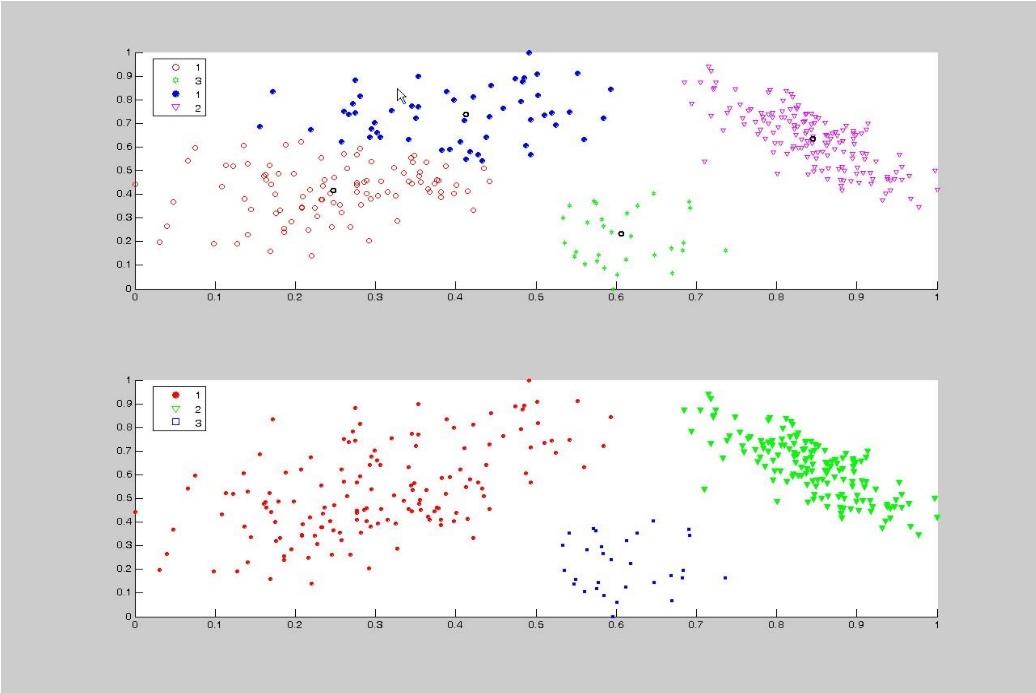

Figure 1 – a) the clustering results of data set set_all;
б) the classification results of data set set_all

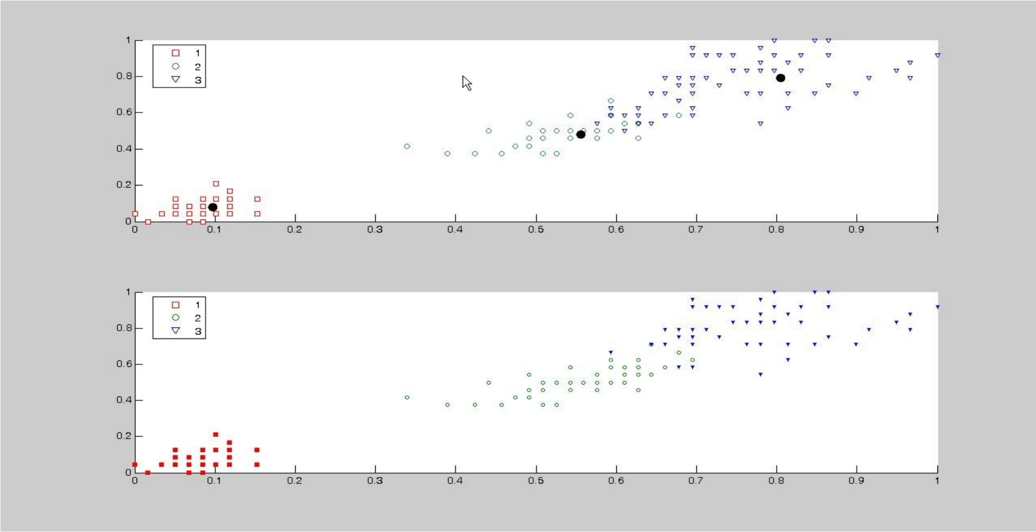

Figure 2 – a) the clustering results of data set Iris (in feature space x_3-x_4);
б) the classification results of data set Iris (in feature space x_3-x_4)