

Stability of Object Classes and Selection of the Latent Features

*A.X.Xurramov**Karshi State University. Karshi, Uzbekistan*Xurramov2009@yandex.ru

Abstract

The problem of selection new latent features space in pattern recognition problem with teacher is considered. To form the space is used by the partition the set of the initial features to the previously unknown number of disjoint groups. The extremum of the criterion of grouping is calculated by the value of stability of each object in class along the set of features included in the group. Displaying such a set to the numerical axis represents the value of the latent feature in the description of admissible objects.

Keywords: Stability of objects, latent features, clustering, generalized estimates, hidden regularities.

Introduction

There are various methods for calculating of the latent (explicitly unmeasurable) features in order to achieve short (compact) description and the convenience of analysis various processes and phenomena. In practice, the factor analysis methods are one of widespread methods for computing of the latent features. These methods have a number of mandatory conditions (restrictions) on their use:

- all features must be quantitative;
- the number of initial features must be twice as many the number of latent;
- the sample must be homogeneous;
- initial features values must be distributed symmetrically;
- covariance matrix between the features are used in calculations.

Using the averaging values in the computation is a significant drawback of factor analysis methods, and narrow the possibilities of its practical application.

Grouping the features is made in order to separate the informative sets among them. In [2] it was provided 5 methods to separate different-type informative features: on the basis of discrete search methods informative zones in the training sample; on the basis of clustering methods; on the basis of the assumption of normal distribution of objects in clusters; on the basis of information theory concepts of entropy; on the basis of nonparametric density estimates. The effectiveness of methods depends on the choice of initial parameters (constraints) and probabilistic assumptions. Theoretically, the number of different subsets of features is limited to $C_n^2 + C_n^3 + \dots + C_n^{n-1}$.

Obviously, the choice of feature space affects to the compactness index.

Development of methods partitioning of the initial set of features into groups in order to form one latent features for description of objects is represented of theoretical and practical interest.

It is proposed to use stability indicators for object classes defines a subset of the original feature set as the data for the grouping criteria.

Selection of the latent features by calculating the indexes of stability is conducted to improve the similarity between the objects belonging to the same class.

The stability of the objects is one of the important concepts in pattern recognition problems. It is one of the indicators for checking the truth of the hypothesis of compactness [1]. On the basis of this hypothesis it

can be develop a criterion for the methods to select the new features space in the description of the objects. The value of stability is the measure of the structural diversity of the sample, which can be used to detect and remove noisy objects in order to improve the generalization ability of the algorithms.

The stability represents local properties of the classified objects. Knowledge of these properties is necessary for determining the anomalous objects of the classes, explaining the reasons of selecting objects as minimal coverage of etalons of the sample, which are enough to correct recognition.

Variety of stability values of the objects class directly depend on choosing of metrics. It was necessary to use different approaches because of closeness measure in different-type features space with metric properties are represented weakly.

In this paper we propose a method of grouping, for which initially did not specify the number and form of group configuration. Partition of initial features into groups is based on the analysis of the paired relationship of closeness and stability of objects.

Under certain dimensions feature space having a "bad" features leads to blurring of relationships of affinity between objects.

To solve this problem it is proposed to use the algorithm of hierarchical agglomerative grouping. The rules of the grouping guarantee the uniqueness of the solution by number and composition of the groups.

We propose to realize:

- partition the initial set of features into disjoint groups by the rule of hierarchical agglomerative grouping and formalize on their base the set of latent features;
- selection of the set of informative latent features;
- detecting and removing the noisy objects to improve the generalizing ability of the recognition algorithms.

1.1 Statement of the problem

We consider the recognition problem in a standard formulation. Objects of the sample belong one of the classes K_1 or K_2 ($E_0 = K_1 \cup K_2$) and each object is described using n different-type features $X(n) = (x_1, \dots, x_n)$, ξ of which are measured in interval scales, and $n - \xi$ is measured on the nominal. We denote the set of indexes of quantitative features through I and of nominal features by J and, here $|I| + |J| = n$.

Let on the sample E_0 the functionality $U(S, X(k))$ is defined to calculate the stability of an object $S \in E_0 \cap K_p$ in class K_p by subset of features $X(k)$, where $X(k), X(k) \subset X(n)$.

The process of building the new features space on a base of partitioning the features to groups is realized in the following order:

- by description of the objects $S_i, i=1, \dots, m, S_i \in E_0$ the set of features $X(n)$ are divided into disjoint groups $X(k_1), \dots, X(k_l), l \leq n, k_1 + \dots + k_l = n$ using the values of $U(S_i, X(k_j)), j=1, \dots, l$;
- the values of features by the group $X(k_j)$ in the description of objects of E_0 are displayed on the numerical axis.

II COMPUTATION OF THE STABILITY OF SAMPLE'S OBJECTS

The concept of the stability of the object naturally follows from the procedure of calculating the optimal neighborhood of the k nearest neighbors, and in this case the stability is understood as a qualitative indicator of the object, describing its location among the objects of the same class. In [Ignomon 1] the stability λ_i^j of object $S_i \in K_j, j = \overline{1, l}, i = \overline{1, m}$ in class K_j is defined as

$$\lambda_i^j = \frac{d_i^j}{2 \max_{1 \leq j \leq l} |K_j| - 3}$$

where d_i^j - the number of events in E_0 , when among the k nearest to S_i objects, $k=1, \dots, 2 \min_{1 \leq j \leq l} |K_j| - 3$ the

majority of them are the objects of $K_j \cap E_0$, obviously, that $\lambda_i^j \in [0, 1]$.

In the paper [ign mad] was presented the way of calculating the stability of object $S_i \in E_0 (S_i = (x_{i1}, x_{i2}, \dots, x_{im}))$ in class K_t , which defined by the formula

$$\gamma_t(s_i) = \frac{1}{n(|E_0 \cap K_t| - 1)} \left(\sum_{c \in I, x_{ic} \in [r_{c_u}, r_{c_v}]} f_{cp}(t)(q_{tp} - 1) + \sum_{c \in J, x_{ic} = p} f_{cp}(t)(h_{tp} - 1) \right) \quad (1)$$

where q_{tp} - the number of values of c -th feature in class K_t in the interval $[r_{c_u}, r_{c_v}]^p$, h_{tp} - number of values of the gradation $x_{ic} = p$ in the description of the object in $E_0 \cap K_t$.

For the algorithm offering in this paper, the measure of similarity between the objects $S_a = (x_{a1}, \dots, x_{an})$ and $S_b = (x_{b1}, \dots, x_{bn})$ of the sample E_0 calculated by the metric of Juravlyev as following:

$$\rho(S_a, S_b) = \sum_{i \in I} |x_{ai} - x_{bi}| + \sum_{j \in J} \begin{cases} 1, & x_{aj} \neq x_{bj}, \\ 0, & x_{aj} = x_{bj}. \end{cases} \quad (2)$$

In (2) the values of the quantitative features are normalized to $[0, 1]$. In order to exclude exhaustive search to find informative features subsets for each object $S_d \in K_p, d=1, \dots, m, p=1, 2$ in the set of $X(k)$ we built no decreasing sequence

$$S_{d_0}, S_{d_1}, \dots, S_{d_{m-1}}, S_{d_m} = S_d \quad (3)$$

by the objects of the sample E_0 , satisfying the inequality $\rho(S_{d_i}, S_d) \leq \rho(S_{d_{i+1}}, S_d), i=0, \dots, m-2$ by the metric (2).

Let we denote as u_p^1, u_p^2 - the number of representatives $K_p, p=1, 2$ in the interval $[c_1^d, c_2^d], (c_2^d, c_3^d]$ accordingly. Here $c_1^d = 0, c_2^d = \rho(S_d, S_{d_h})$ and $c_3^d = \rho(S_d, S_{d_{m-1}})$, h - order number in the sequence (3).

The definition of the interval boundary c_2 is based on the criterion about that, each interval $[c_1, c_2], (c_2, c_3]$ is consisted the values of distance by the metric (2) of the representatives of only one class and its extreme value by the subset $X(k)$ is calculated as following:

$$\left(\frac{\sum_{i=1}^2 u_i^1 (u_i^1 - 1) + u_i^2 (u_i^2 - 1)}{\sum_{i=1}^2 |K_i| (|K_i| - 1)} \right) \left(\frac{\sum_{d=1}^2 \sum_{i=1}^2 u_i^d (|K_{3-i}| - u_{3-i}^d)}{2|K_1||K_2|} \right) \rightarrow \max_{c_1 < c_2 < c_3} \quad (4)$$

For the intervals building by the criterion (4), we define $\alpha_1(p) = \left| \left\{ S_{d_i} \in K_p \mid \rho(S_d, S_{d_i}) \in [c_1, c_2] \right\} \right|$, $\alpha_2(p) = \left| \left\{ S_{d_i} \in K_{3-p} \mid \rho(S_d, S_{d_i}) \in [c_1, c_2] \right\} \right|$, $\theta_1(p) = \alpha_1(p) / |K_p|$ и $\theta_2(p) = \alpha_2(p) / |K_{3-p}|$. The stability of the object $S_d \in K_p, p=1, 2$ by subset of the features $X(k)$ is calculated as following:

$$U(S_d, X(k)) = \theta_1(p)(1 - \theta_2(p)). \quad (5)$$

III COMPUTATION OF THE GENERALIZED ESTIMATIONS OF SAMPLE'S OBJECTS

The generalized estimation is a value of combined indices, obtained on the base of numerical analysis of opposition like - «*patient – practically healthy*», «*hard worker-loafer*», «*rich-poor*», «*excellent student-debtor student*», «*generous- greedy*» and so on.

The data about oppositions may be presented as the sample objects of two disjoint classes at the formalization of the problem. The meaning of computing the estimation leads to explain the process of decision, for example, patients diagnosis by the value of the degree of disease severity.

Computation of generalized evaluations occurs by means aggregation, as a rule, the initial features. The results of aggregation may considered as values of the new latent features in the description of permissible objects. The generalized estimation may be used for displaying the relationship between objects of two classes in different-type features space on the numerical axis.

To calculate the estimation of any permissible object $S = (x_1, \dots, x_n)$ by the subset of features $X(k)$ is used the functional:

$$R(S) = \sum_{i \in X(k)}^n w_i t_i (x_i - c_1^i) / (c_2^i - c_0^i) \quad (6)$$

where c_0^i, c_1^i and c_2^i are the boundaries of the interval by (3) of ordered sequence of i -th feature, $t_i \in \{-1, 1\}$, and the vector $T = (t_1, \dots, t_n)$ is defined by following condition:

$$\min_{S_p \in K_1} R(S_p) - \max_{S_p \in K_2} R(S_p) \rightarrow \max . \quad (7)$$

Searching the solution of the multiextremal problems by (7) is realized using the stochastic optimization algorithm. Step by step implementation of this algorithm is as follows:

1. Selection the number of iterations $k, \left\lceil \frac{n}{2} \right\rceil \leq k \leq n$, $iter = 0$, $T_{\max} = \overbrace{(1, \dots, 1)}^n$, $Z_{\max} = -n$;
2. $iter = iter + 1$. $\Omega = \{1, \dots, n\}$. Selection the initial values of the vector $T = (t_1, \dots, t_n)$ for new $iter$ -th iteration. Computation the value $R(S_j) (S_j = (x_{j1}, \dots, x_{jn})) \forall S_j \in E_0$ and $Z = \min_{S_j \in K_1} R(S_j) - \max_{S_j \in K_2} R(S_j)$;
3. $\forall i \in \Omega$ calculation the value (under the condition replacing t_i to $-t_i$)
 $Z_i = \min_{S_j \in K_1} R^*(S_j) - \max_{S_j \in K_2} R^*(S_j)$
 where $R^*(S_j) = R(S_j) - 2t_i w_i (x_{ji} - c_1^i) / (c_2^i - c_0^i), j = \overline{1, m}$;
4. $Z_p = \max_{i \in \Omega} Z_i$. If $Z_p > Z$, to $Z = Z_p$, $\Omega = \Omega \setminus p$,
 $R(S_j) = R(S_j) - 2t_p w_p (x_{jp} - c_1^p) / (c_2^p - c_0^p), j = \overline{1, m}, t_p = -t_p$,
 and go to 3;
5. If $Z > Z_{\max}$ then $Z_{\max} = Z$, $T_{\max} = T$;
6. If $iter < k$ then 2;
7. Output Z_{\max} , T_{\max} .

The steps of the algorithm from 2 to 4 represent the calculation of local maximum at the different initial values of elements of the vector T . The maximum value of Z_{\max} among the local maxima and the corresponding values of the vector elements T_{\max} (Step 5) is selected as solution of the problem on the condition (7).

For objects with the description in different-type features space additionally require to define the

weights of nominal features and their gradations too.

We denote by p the number of gradation of the feature $r \in J$, g_{dr}^t - number of values of t -th ($1 \leq t \leq p$) gradation of r -th feature in the description of the objects of class K_d , l_{dr} - the number of gradation of r -th feature in K_d .

The difference on the r -th feature between classes K_1 and K_2 is defined as the value of

$$\lambda_r = 1 - \frac{\sum_{t=1}^p g_{1r}^t g_{2r}^t}{|K_1||K_2|}. \quad (8)$$

A value of β_r - the degree of uniformity of classes K_1 and K_2 on the r -th feature (similarity measure in class) is calculated according to the formulas:

$$D_{dr} = \begin{cases} (|K_d| - l_{dr} + 1)(|K_d| - l_{dr}), & p > 2, \\ |K_d|(|K_d| - 1), & p \leq 2; \end{cases}$$

$$\beta_r = \begin{cases} \frac{\sum_{t=1}^p g_{1r}^t (g_{1r}^t - 1) + g_{2r}^t (g_{2r}^t - 1)}{D_{1r} + D_{2r}}, & D_{1r} + D_{2r} > 0, \\ 0, & D_{1r} + D_{2r} = 0. \end{cases} \quad (9)$$

Using the formulas (8), (9) the weight of nominal feature $r \in J$ is determined as

$$v_r = \lambda_r \beta_r. \quad (10)$$

It is easy to verify that the set of weights values of nominal and quantitative features, calculated by (4) and (10) belong into the interval $[0,1]$.

It is obvious that the set of numbers identifying as p gradation of the nominal feature, it is always possible one-to-one mapping into the set $\{1, \dots, p\}$. Taking into account such mapping for the object $S = (x_1, \dots, x_n)$ contribution of the feature $x_i = j$, $i \in J$, $j \in \{1, \dots, p\}$ in generalized estimation is determined by the

$$\mu_i(j) = v_i \left(\frac{\alpha_{ij}^1}{|K_1|} - \frac{\alpha_{ij}^2}{|K_2|} \right) \quad (11)$$

where $\alpha_{ij}^1, \alpha_{ij}^2$ - are the numbers of the values the j -th gradation of the i -th feature accordingly in classes K_1 and K_2 , v_i - is a weight of the i -th feature, calculating by (10).

IV THE FEATURES GROUPING ALGORITHM

Agglomerative hierarchical algorithm is used the similarity matrix values $B = \{b_{ij}\}$ when selecting the candidates for sequential grouping the features.

The distance between the objects $S_a, S_b \in E_0$ by the features $\{x_i, x_j\} \subset X(n)$ is calculated as follows:

$$\beta(i, j, S_a, S_b) = \begin{cases} \rho(S_a, S_b), & i \neq j, \\ 0, & \text{otherwise.} \end{cases}$$

The b_{ij} element of the similarity matrix $B = \{b_{ij}\}_{n \times n}$ is a measure of the contribution of a pair of features $\{x_i, x_j\}$ in division of the sample E_0 into two classes K_1, K_2 and is determined as

$$b_{ij} = \begin{cases} \frac{\sum_{a=1}^m \sum_{b=1}^m \alpha(S_a, S_b) \beta(i, j, S_a, S_b)}{\sum_{p=1}^2 |K_p| (m - |K_p|)}, & i \neq j, \\ 0, & \text{otherwise,} \end{cases} \quad \text{where } \alpha(S_a, S_b) = \begin{cases} 0, & S_a, S_b \in K_i, i = 1, 2, \\ 1, & S_a \in K_i, S_b \in K_{3-i}. \end{cases}$$

The rule of hierarchical agglomerative grouping is based to analyze the relationship of objects stability by (5) on a set of initial features. The set $X(k)$ is considered to be formed, if the inclusion of another feature lowers the value of stability of the objects.

Changing the stability of objects of the sample E_0 on the description from $X(k)$ to $X(k+1), 1 \leq k < n$, $X(k) \subset X(k+1)$ is calculated as follows:

$$M(X(k), X(k+1)) = \frac{1}{m} \sum_{d=1}^m \begin{cases} 1, & U(S_d, X(k)) \leq U(S_d, X(k+1)), \\ 0, & \text{otherwise.} \end{cases}$$

The features grouping algorithm consists of the following steps:

Step 1. $l = 0, \Omega = \{1, \dots, n\}$;

Step 2. If $\Omega = \emptyset$, then go to Step 6, else $l = l + 1, G_l = \emptyset$;

Step 3. $b_{ij} = \max_{u, v \in \Omega} \{b_{uv}\}_{n \times n}$, $G_l = \{i, j\}$, $\Omega = \Omega \setminus \{i, j\}$;

Step 4. If $\Omega = \emptyset$, then go to Step 6, else $Mon(p) = \max_{p \in \Omega} \{M(G_l, G_l \cup \{p\})\}$;

Step 5. If $Mon(p) \geq 0.5$, then $G_l = G_l \cup \{p\}$, $\Omega = \Omega \setminus \{p\}$ and go to Step 4, else go to Step 2;

Step 6. End.

The generalized estimation $R_k(S_a)$ (a latent feature) [2] of the object $S_a = (x_{a1}, \dots, x_{am}) \in E_0, a = 1, \dots, m$ by the group $G_k (1 \leq k \leq l)$ - is calculated as

$$R_k(S_a) = \sum_{i \in G_k \cap I} w_i t_i (x_{ai} - c_2^i) / (c_3^i - c_1^i) + \sum_{i \in G_k \cap J} \mu_i (x_{ai}). \quad (12)$$

Let we denote through $w(G_k)$ the optimal value of the criterion (4) by the sequence

$$R_k(S_1), \dots, R_k(S_m).$$

It is common knowledge that existence of the property of monotonicity on the calculation extends the possibility of using the standart recognition algorithms for obtained results. Calculating the generalized estimations by formula (12), the sum of the quantitative features are depended on the random values t_i , the monotonicity property in the sequence of construction the groups is satisfied only on condition $G_k \cap I = \emptyset (1 \leq k \leq l)$ and defined as

$$w(G_p) \geq w(G_q), p < q.$$

V. DETECTION AND REMOVAL OF THE «NOISE» OBJECTS

The presence in the sample poorly classified objects called as “«noise» object” negatively affects the quality of the results of recognition algorithms. Therefore, in many cases preprocessing the sample is implemented in order to remove the «noise» objects. In the presence of «noise» objects in the sample the

various algorithms of recognition process them differently. The algorithms for constructing decision trees to reduce the impact of such objects is provided the reduction procedure – remove the subtrees having a low statistical reliability due to the fact that their construction used object-emissions. In other algorithms data preprocessing is provided during it, the «noise» objects are detected and removed using some criterion. In some cases, even an attempt to adjust the individual features of the object-emission with the aim to transform it into a typical object. Most of the procedures are based on using the “nearest neighbor” rules to detect the «noise» objects

In [3] is considered the problem of censoring the sample, initially containing a significant number of incorrectly classified objects. Offered the algorithm of censoring which oriented only on local characteristics of the sample objects. To assess the probability of the object belonging to one of the two images used ternary relative measure - the function of rival similarity (FriS function). In the fixed features space the censoring is consists with the sequential removal of objects, most degrade the quality sample descriptions (or separability estimation of classes). The results of testing the algorithm on a wide range of model problems led to the conclusion that the objects that were deleted before the function inflexion point describing the separability of classes, as a rule, emissions, distorting the structure of the data.

It is offered the following way searching the «noise» objects. Let for $S_k \in K_i, \rho(S_k, S_r) = \min_{S_j \in CK_i} \rho(S_k, S_j)$ and $Z = \left| \left\{ S_\mu \in K_i \mid \rho(S_k, S_\mu) < \rho(S_k, S_r) \right\} \right|$. We denote through $D_i (D_i \in CK_i)$ the set of «noise» object of the class K_i . Object $S_r \in CK_i$ is included into D_i and considered as noisy if performed the following condition

$$\frac{ZZ - \lambda}{|K_i|} > \frac{1}{m - |K_i|} \quad (13)$$

where $ZZ = \left| \left\{ S_\mu \in K_i \mid \rho(S_r, S_k) < \rho(S_\mu, S_k) < \rho(S_\eta, S_k) \right\} \right|, |\lambda| \geq 0, \rho(S_\eta, S_k) = \min_{S_j \in CK_i \setminus \{S_r\}} \rho(S_j, S_k)$. The value Z and $Z + ZZ$ may be considered as the number of representatives of the class K_i in the hypersphere with the center at $S_k \in K_i$ accordingly before and after removal the «noise» object S_r .

VI. COMPUTATIONAL EXPERIMENT

The computational experiment was conducted with data of Echocardiogram [4]. In reality, the sample consists of 132 objects, but in this experiment, we did not consider the case of missing values in data. Therefore, all objects which have missing values are removed. The sample consists of 108 objects, each object is described by 9 quantitative and 2 nominal features. The sample contained 74 representatives of the class K_1 (patient was either dead after 1 year or had been followed for less than 1 year) and 34 of the class K_2 (patient was alive at 1 year).

In table 1 is presented the results of the division the features set into disjoint groups by using features grouping algorithm described above.

Table 1. Division the features into disjoint groups.

The group number (latent variable)	The group and its elements (feature number)	The value of the criterion (4)
1	$X(1) = \{1, 2\}$	0.748106
2	$X(2) = \{3, 4, 5, 6, 7, 8, 9, 10, 11\}$	0.347756

The value of the criterion (4) is calculated using (6) on the set of features $X(1)$ and $X(2)$.

In order to increase the generalizing ability of the algorithm is performed the detection the «noise» objects by (13). In this goal, we can detect the «noise» objects with some value of the heuristic parameter λ . The full information about using the method of detecting the «noise» objects with parameter λ you can find in [4]. There are presented the indexes of «noise» objects when $\lambda=0$ in table 2.

Table 2. The indexes of the «noise» objects by class.

№	Class	The number of «noise» object
1	K_1	1,2,7,26,34,38,48,58,64,83,87,91
2	K_2	5,17,37,50,69,71,78,97,104

The effect of removing the «noise» objects (see table 2) is demonstrated in table 3.

Table 3. Division the features into disjoint groups when removing the «noise» objects.

The group number (latent variable)	The group and its elements (feature number)	The value of the criterion (4)
1	$X(1)=\{1,2,8\}$	0.949484
2	$X(2)=\{3,4,6,7,9,10,11\}$	0.414243
3	$X(3)=\{5\}$	0.302811

Table 3 shows that the number of groups increased by one and consists of 3 groups. And the value of separability by criterion (4) is increased by the first and second latent features than initial (see 1, 2 rows of table 1 and table 3).

For visualization the objects of the sample selected the first two latent features obtained before removal (see table 1) and after removing the «noise» objects (see table 3). The change in the structure of class objects can be traced in Fig. 1 and Fig. 2

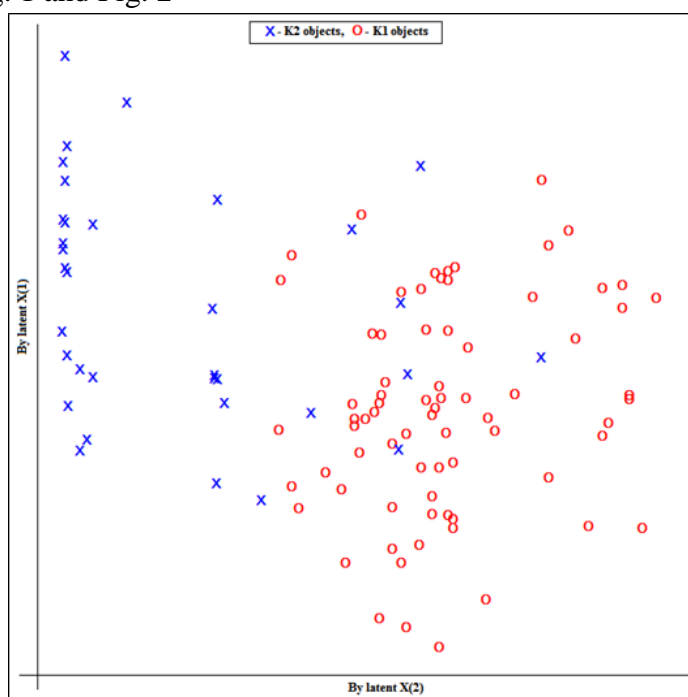


Figure 1. The structure of objects of classes without removing the «noise» objects.

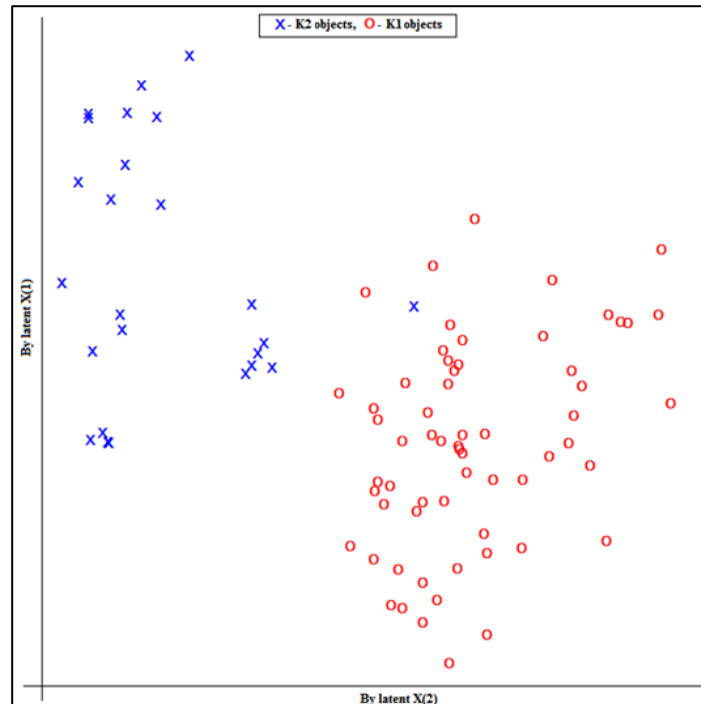


Figure 2. The structure of objects of classes after removing the «noise» objects.

In Fig. 1 and Fig. 2, the objects of class K_1 are marked with the symbol "o", K_2 - "x".

By results of the computational experiment (see. Table. 1, 2, 3,4) may be done the following conclusions:

- agglomerative hierarchical grouping algorithm allows to order the latent features by informativeness;
- removal the «noise» objects increases the stability of the objects of the sample by latent features (12) on the defined set of the features.

REFERENCES

- [1] Ignatev N.A. “Obobshennie otsenki i lokalnie metriki ob’ektov v intellektualnom analize dannix”, Natsionalniy universitet Uzbekistana im. Mirzo Ulugbeka, Tashkent, 72 p, 2014.
- [2] Kolesnikova S.I. “Metodi analiza informativnosti raznotipnix priznakov”, Vestnik tomskogo gosudarstvennogo universiteta, Tomsk, №1(6), pp. 69-80, 2009.
- [3] Borisova I.A., Kutnenko O.A “Senzurirovaniye oshibochno klassifitsirovannykh ob’ektov viborki”, Mashinnoe obuchenie i analiz dannykh, tom: 1, №11, pp. 1632-1641, 2015.
- [4] Ignatyev N.A. “Intellektualniy analiz dannix I gipoteza o kompaktnosti klassov. Meri kompaktnosti, kriterii otsenok”, Palmarium Academic Publishing, Saarbrücken, 92 p., 2016.
- [5] <http://archive.ics.uci.edu/ml/machine-learning-databases/echocardiogram>