I.J. Intelligent Systems and Applications, **2017, 8, 51-58**
Published Online August 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2017.08.06

Data Visualization and its Proof by Compactness Criterion of Objects of Classes

Saidov Doniyor Yusupovich
Algorithms and programming technology, National University of Uzbekistan named after
Mirzo Ulugbek, Tashkent, 100174, Uzbekistan
E-mail: doniyor_2286@mail.ru

Received: 21 October 2016; Accepted: 05 June 2017; Published: 08 August 2017

Abstract—In this paper considered the problem of reducing the dimension of the feature space using nonlinear mapping the object description on numerical axis. To reduce the dimensionality of space used by rules agglomerative hierarchical grouping of different - type (nominal and quantitative) features. Groups do not intersect with each other and their number is unknown in advance. The elements of each group are mapped on the numerical axis to form a latent feature. The set of latent features would be sorted by the informativeness in the process of hierarchical grouping. A visual representation of objects obtained by this set or subset is used as a tool for extracting hidden regularities in the databases. The criterion for evaluating the compactness of the class objects is based on analyzing the structure of their connectivity. For the analysis used an algorithm partitioning into disjoint classes the representatives of the group on defining subsets of boundary objects. The execution of algorithm provides uniqueness of the number of groups and their member objects in it.

The uniqueness property is used to calculate the compactness measure of the training samples. The value of compactness is measured with dimensionless quantities in the interval of [0, 1]. There is a need to apply of dimensionless quantities for estimating the structure of feature space. Such a need exists at comparing the different metrics, normalization methods and data transformation, selection and removing the noise objects.

Index Terms—Data visualization, logical regularities, nonlinear mapping, compactness of objects.

I. Introduction

The visualization is a powerful tool for searching hidden regularities at implementations of data mining techniques. In most well-known visualization techniques applied a linear mapping of the original space into one, two and three-dimensional space. The main goal pursued by researchers consists of analyzing the relationship between objects. As a rule, such an analysis makes it possible to detect hidden patterns in the data that represent the new knowledge in one or another subject area.

One of the causes of using visualization methods in data mining is to find logical patterns. The complexity of this searching is related to the decision of the NP - complete problem. A classic example of using the visualization in data mining is the method of the local geometry [1]. Searching logical regularities by using this method required an active participation of researchers in the process. New knowledge obtained from the database by applying statistical methods to them on the results of the visualization.

In [1] it was concluded that for the research of structure of the set of logical rules most suitable methods are multidimensional scaling and hierarchical agglomerative procedures of cluster analysis. These methods allow to get the visual representation of the geometric structure of the aggregate logical regularities and their results are complemented each other.

In this paper we offer to apply a nonlinear mapping the description of objects on numerical axis with the rules of hierarchical agglomerative grouping. It is proved the truth of the statement that latent features which obtained as a result of mapping express the compactness of class objects better than the initial space of the raw features. The latent features which formed as a result of grouping, obtained ordered by the degree of informativeness.

In this paper considered the criteria for the estimation of different visualization methods. Most often visualization is interpreted as a problem of reducing the dimension of the space by the transition to the new features in the description of objects. There are a lot of criterions which optimizations [2] aim to preserve the original and the new space structure through the distance between objects. The estimation of local compactness is oriented on checking the similarity of class objects (including the visualisation) [3].

In this paper presented the basis of using the criterion for analyzing the structure of relatedness of objects through the so-called shell (a subset of the boundary objects) classes. As the shell [4] used a subset of the boundary objects of classes by a given metric. The value of the offering criterion determines the degree of compactness of classes in [0,1].

The number of features plays a significant role in structured datasets. We can apply the standard algorithms and methods without difficulty to analyze the data if the size of feature space rather small, otherwise there may be a major problem, known as “curse of dimensionality”, that reduces the classification accuracy, increases the classification model complexity and increases the computational time, thus the need for Feature Selection (FS) [5] [6]. In [7] authors offered the feature selection approaches, which are capable of minimizing the number of selected features without reducing the classification accuracy of all features .

Searching informative features set usually be used to estimate the classification accuracy of the algorithms. In [8] a genetic algorithm (GA)-based features selection to improve the accuracy of medical data classification is proposed. In [9] developed a novel feature selection technique based on the Partial Least Squares (PLS). PLS aims to obtain a low dimensional approximation of a matrix that is ‘as close as possible’ to a given vector. The main purpose of the proposed method is to select the significant feature subset which gives the higher classification accuracy with the different classifiers.

In the form of the computational experiment on the test example proved that the degree of compactness in nonlinear data mapping higher than the method of PCA.

This paper is organized as follows. Section I provides an overview of the most significant related work and the structure of the paper. Section II describes the problem statement, research purpose, requirements and theoretical basis. In section III discussed some another method to verify the obtained result and an algorithm based on this method. Section IV presents the computational experiment. Section V concludes the paper.

II. Problem Statement

The two class recognition problem is considered in a standard formulation. Objects of the sample $E_0 = \{S_1, \ldots, S_m\}$ belong to one of the classes K_1 or K_2 $(E_0 = K_1 \bigcup K_2)$ and each object is described using n different – type features $\mathrm{X(n)} = (\mathrm{x}_1, \ldots, \mathrm{x}_n)$, a set of permissible values ξ of which is measured on the interval scales, and $n-\xi$ is measured on the nominal scale. It is considered that in $X(n)$ given a linear and nonlinear mapping of objects descriptions on numerical axis. The nonlinear mapping is implemented by sequential partitioning a set of $X(n)$ according to the rule of hierarchical agglomerative groupings in $X_1(k_1), \ldots, X_\tau(k_\tau), \tau \geq 1, k_1 + \ldots + k_\tau \leq n$ disjoint subsets. Each subset of $X_1(k_t), t = 1, \ldots, \tau$ is mapped on the numerical axis and is considered as a new latent feature in the description of the objects. It is required:

- generate a new space of latent features for mapping object descriptions to the plane with the algorithms of the linear and nonlinear methods;
- give a numerical estimation of placement structure of the class object descriptions in space which based on defined measure proximities.

To implement linear and nonlinear methods used proof of the truth of the compactness hypothesis in the form of a computational experiment. The proof is based on the following criteria which proposed partitioning the values of feature into intervals.

Let us assume that on E_0 defined the ordering of objects by any of its property q , which given (initial values of a quantitative attribute), or defined by an algorithmic way.

$$q(S_{i_1}),\ q(S_{i_2}),\ \ldots,\ q(S_{i_m}). \qquad (1)$$

The ordered set of values (1) is divided into two disjoint intervals $[c_0, c_1]$, $(c_1, c_2]$, each of them is considered as the gradation of the nominal feature. The criterion for determining the boundaries of c_1 is based on checking the hypothesis (statement) that each of the two intervals contains the values of a quantitative value of only one class objects.

Let u_i^1, u_i^2 - the number of values (1) by some quantitative (the initial or latent) feature of class $K_i, i = 1, 2$ accordingly in the intervals $[c_0, c_1]$, $(c_1, c_2]$, $|K_i| > 1$, v - the sequential number of the ordered element in ascending order (1) in E_0 , which defines the boundaries of intervals as $c_0 = q(S_{i_1}),\ c_1 = q(S_{i_v}),\ c_2 = q(S_{i_m})$. The criterion

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

allows to calculate the optimal value of the boundaries between intervals $[c_0, c_1], (c_1, c_2]$. The expression on the left-hand brackets (2) is the intraclass similarity and in the right - interclass difference.

For the convenience of exposition, we denote the set of indexes of quantitative features through I , and nominal features by J . We consider the calculation of the weights and contribution of the nominal features grades which are used in both linear and nonlinear methods.

Let us denote by p the number of gradation of the feature $r \in J$, g_{dr}^t - a number of values of the t -th ($1 \leq t \leq p$) gradation of the r -th feature in the description of objects of the class K_d , l_{dr} - a number of gradation of the r -th feature in K_d . The difference over the r -th feature between the classes K_1 and K_2 is determined as a value

$$\lambda_r = 1 - \frac{\sum_{t=1}^{p} g_{1r}^t g_{2r}^t}{|K_1| |K_2|} . \qquad (3)$$

The degree of uniformity (intraclass similarity measure) β_r gradation values of the r-th feature by classes of K_1, K_2 is calculated according to the formulas:

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

$$\beta_r = \begin{cases} \dfrac{\sum\limits_{t=1}^{p} g_{1r}^t\left(g_{1r}^t - 1\right) + g_{2r}^t\left(g_{2r}^t - 1\right)}{D_{1r} + D_{2r}}, D_{1r} + D_{2r} > 0, \\ 0, \qquad D_{1r} + D_{2r} = 0. \end{cases} \quad (4)$$

With (3), (4) the weight of nominal feature $r \in J$ is determined as

$$v_r = \lambda_r \beta_r . \quad (5)$$

It is easy to verify that the set of weight values of nominal and quantitative features, calculated by (2) and (5) belong to the interval $[0,1]$.

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

$$\mu_i\left(j\right) = v_i\left(\frac{\alpha_{ij}^1}{|K_1|} - \frac{\alpha_{ij}^2}{|K_2|}\right) \quad (6)$$

where $\alpha_{ij}^1, \alpha_{ij}^2$ -are the numbers of the values of j th gradation by i-th feature accordingly in classes K_1 and K_2, v_i - is a weight of the i-th feature, calculating by (5). In different – type features space generalized estimation for each object $S_a \in E_0$, $S_a = \left(x_{a1},..., x_{an}\right)$ will be calculated as

$$R\left(S_a\right) = \sum_{i\in I} w_i t_i \left(x_{ai} - c_1^i\right) / \left(c_2^i - c_0^i\right) + \sum_{i\in J} \mu_i\left(x_{ai}\right) \quad (7)$$

where $\mathrm{w}_i, i \in \mathrm{I}$, and the boundary of the interval c_1^i is calculated as optimal value of the feature $x_i \in X(n)$ by (2) on E_0 , and the vector $T = \left(t_1,..., t_n\right), t_i \in \{-1,1\}$ is defined from the condition

$$\min_{S_a \in K_1} R\left(S_a\right) - \max_{S_a \in K_2} R\left(S_a\right) \to \max_{E_0} . \quad (8)$$

In presenting objects by linear method on the plane will select two numerical scales for mapping to them descriptions of objects with features in $X(n)$ using functional R_1, R_2 . In E_0 defined division into two disjoint subsets E_0^1, E_0^2 of the representatives of K_1, K_2. In E_0^1, E_0^2 calculated the optimal values of the functional R_1 , R_2 according to (8), which are used for mapping objects of E_0 in two numerical scales. It is proved the equivalence of these two scales by (2) in [10].

In algorithms of nonlinear mapping to calculate the values of defined set of features in the description of objects used hierarchical agglomerative grouping rules [11]. For identifying the features as an initial so an obtained by calculating the generalized estimation on the p-th step $(0 \leq \mathrm{p} < n)$ of grouping will use $\left\{\mathrm{x}_\mathrm{i}^\mathrm{p}\right\}_{i\in I\cup J}$, where $|I| + |J| = n$ at $p = 0$. The extremum of the criterion (2) is used as a weight $w_j^p \left(0 \leq w_j^p \leq 1\right)$ of the feature x_j^p . At $w_j^p = 1$ the values of the feature x_j^p of objects in classes K_1, K_2 do not intersect between themselves. The value of generalized estimation b_{rij}^p (latent feature's) of the object $S_r = (a_{r1}^p, \ldots, a_{r,n-p}^p)$, $S_r \in E_0$ by pair of $\left(x_i^p, x_j^p\right)$, $1 \leq p < n, ?i, j \in I, ?i \neq j$ is calculated as

$$b_{rij}^p = \begin{cases} \mu_i(\mathrm{a}_{ri}^p) + \mu_j(\mathrm{a}_{rj}^p), \text{ i, j} \in \mathrm{J} \\ \mu_i(\mathrm{a}_{ri}^p) + t_j w_j^p \left(\dfrac{a_{rj}^p - c_2^{jp}}{c_3^{jp} - c_1^{jp}}\right), i \in J, j \in I \\ \eta_{ij}\left(t_i w_i^p \left(\dfrac{a_{ri}^p - c_2^{ip}}{c_3^{ip} - c_1^{ip}}\right) + t_j w_j^p \left(\dfrac{a_{rj}^p - c_2^{jp}}{c_3^{jp} - c_1^{jp}}\right)\right) + \\ + \left(1 - \eta_{ij}\right) t_{ij} w_{ij}^p \left(\dfrac{a_{ri}^p a_{rj}^p - c_2^{ijp}}{c_3^{ijp} - c_1^{ijp}}\right), i, j \in I \end{cases} \quad (9)$$

where $\mu_i(\mathrm{a}_{ri}^p)$, $\mu_j(\mathrm{a}_{rj}^p)$ are calculated by (6), $w_i^p, \mathrm{w}_j^p, \mathrm{w}_{ij}^p$ – are the weights of the features, which defined by (2) accordingly by the set of features values x_i^p, x_j^p and their multiplication $x_i^p x_j^p$, the values $t_i, t_j, t_{ij} \in \{-1,1\}$, $\eta_{ij} \in [0,1]$ are selected by extremum of the functional

$$\varphi\left(p, i, j\right) = \frac{\min\limits_{S_r \in K_1} b_{rij}^p - \max\limits_{S_r \in K_2} b_{rij}^p}{\max\limits_{S_r \in E_0} b_{rij}^p - \min\limits_{S_r \in E_0} b_{rij}^p} \to \max . \quad (10)$$

The value (10) is interpreted as an offset between objects of the classes K_1 and K_2. A detailed description of the algorithm for selecting the latent features by (9) and (10) can be found in [11].

The latent features calculated by the (9) form a new space for the object descriptions in recognition algorithms. Analytical representation of nonlinear transformations proposed in [12], makes it possible to use the results of the described algorithm as a decision rule for recognition.

In nonlinear methods opposed to the linear, we cannot say about the equivalence of two numerical scales for the display of objects descriptions. It is proposed to use the values of the first two latent features for visualization because of they have the highest indexes of separability of objects projection by criterion (2). The proof of the statement is based on the criterion of compactness of class objects.

III. Cluster Analysis and Compactness of Class Objects

The main purpose of dividing the class objects into disjoint groups is the calculation and analyzing compactness values of objects of classes and sample as a whole by the result of visualization. The compactness is offered to calculate by the results of grouping algorithm [13] of representatives of classes according to their relatedness through the defined subset of the boundary objects. As a distance $\rho(\mathrm{x},\mathrm{y})$ between the objects in E_0 is used a metric of Juravlyev. The metric of Juravlyev allows analyzing the structure of class objects considering the features variety.

Let we denote through $L(\mathrm{E}_0,\rho)$ - the subset of boundary objects of classes, which is defined in E_0 by the metric $\rho(\mathrm{x},\mathrm{y})$. The objects $S_i,S_j \in K_t, t=1,2$ are considered to be related to each other $(\mathrm{S}_i \leftrightarrow \mathrm{S}_j)$, if $\{\mathrm{S} \in \mathrm{L}(\mathrm{E}_0,\rho) \mid \rho(\mathrm{S},\mathrm{S}_i) < \mathrm{r}_i \text{ and } \rho(\mathrm{S},\mathrm{S}_j) < \mathrm{r}_j\} \neq \varnothing$, where $r_i(r_j)$ - is the distance from $S_i(S_j)$ to the nearest object in K_{3-t} by the metric $\rho(\mathrm{x},\mathrm{y})$.

The set $G_{tv}=\{\mathrm{S}_{v_1},...,\mathrm{S}_{v_c}\}$, $\mathrm{c} \geq 2$, $G_{tv} \subset K_t, v \leq |K_t|$ presents the area (group) with related objects in class K_t, if for any $\mathrm{S}_{v_1},\mathrm{S}_{v_c} \in G_{tv}$ there is a way $\mathrm{S}_{v_i} \leftrightarrow \mathrm{S}_{v_k} \leftrightarrow ... \leftrightarrow \mathrm{S}_{v_j}$. Object $S_i \in K_t, t=1,2$ belong to the group with single element and is considered not related, if there is no way $S_i \leftrightarrow S_j$ for any object $S_i \neq S_j$ and $S_j \in K_t$. It is required to define the minimum number of disjoint groups of related and unrelated objects for each class $K_t, t=1,2$.

To define the minimum number of groups of related and unrelated class objects we use $L(\mathrm{E}_0,\rho)$ - the subset of the boundary objects (shell) of classes by metric Juravlyev ρ and it allows us to describe the objects in the new space of binary features.

To separate the shell (boundary objects) of the class for each object $S_i \in K_t, t=1,2$ would built an ordered sequence by $\rho(\mathrm{x},\mathrm{y})$ as

$$S_{i_0}, S_{i_1}, ..., S_{i_{m-1}}, S_i = S_{i_0}. \tag{11}$$

Let $S_{i_\beta} \in K_{3-t}$ is the nearest to S_i object from (11) not including in class K_t. Let us denote through $O(\mathrm{S}_i)$ the neighborhood of S_i, with radius $r_i = \rho(\mathrm{S}_i,\mathrm{S}_{i_\beta})$ and centered on S_i, including all of the objects for which $\rho(\mathrm{S}_i,\mathrm{S}_{i_\tau}) < \mathrm{r}_i, \tau = 1,...,\beta-1$. In $O(\mathrm{S}_i)$ there is always nonempty subset of objects

$$\Delta_i = \{\mathrm{S}_{i_\alpha} \in \mathrm{O}(\mathrm{S}_i) \mid \rho(\mathrm{S}_{i_\beta}, \mathrm{S}_{i_\alpha}) = \min_{\mathrm{S}_{i_\tau} \to \mathrm{O}(\mathrm{S}_i)} \rho(\mathrm{S}_{i_\beta}, \mathrm{S}_{i_\tau})\}. \tag{12}$$

By (12) the belonging of objects to the shell of classes is defined as $L(\mathrm{E}_0,\rho) = \bigcup_{i=1}^{m} \Delta_i$.

The set of shell objects in $K_t \cap L(\mathrm{E}_0,\rho)$ is denoted as $L_t(\mathrm{E}_0,\rho) = \{\mathrm{S}^1,...,\mathrm{S}^\pi\}, \pi \geq 1.$ The value $\pi = 1$ unambiguously determines including all of the objects of classes into one area. When $\pi \geq 2$ we transform the description of each object $S_i \in K_t$ to $S_i = (\mathrm{y}_{i1},...,\mathrm{y}_{i\pi})$, where

$$y_{ij} = \begin{cases} 1, \rho(\mathrm{S}_i,\mathrm{S}^j) < \mathrm{r}_i, \\ 0, \rho(\mathrm{S}_i,\mathrm{S}^j) \geq \mathrm{r}_i. \end{cases} \tag{13}$$

Let by (13) the description of objects of class K_t is obtained in new (binary) feature space, $\Omega = K_t, \theta$ - number of disjoint groups of objects, and $S_\mu \wedge S_\eta \in K_t$. The step by step execution of the algorithm of partitioning the objects of K_t into disjoint groups $G_1,...,G_\theta$ is given as follows:

Step 1: $\theta = 0$;

Step 2: Separate the object $S \in \Omega$, $\theta = \theta + 1$, $Z = S$, $G_\theta = \varnothing$;

Step 3: Perform the selection $S \in \Omega$ *and* $S \wedge Z = true$, $\Omega = \Omega \setminus S$, $G_\theta = G_\theta \cup S$, $Z = Z \vee S$ until $\{\mathrm{S} \in \Omega \mid S \wedge Z = true\} \neq \varnothing$;

Step 4: If $\Omega \neq \varnothing$, then go to 2;

Step 5: End.

To analyze the results of visualization proposed to use the structural characteristic as an estimation of the compactness of classes

$$\theta_i = \frac{\sum_{j=1}^{\mu} m_{ij}^2}{m_i^2} \tag{14}$$

where μ - is the number of groups in $K_i, i=1,2$, $m_i = |K_i|$, m_{ij} - is the number of objects in the j-th group of the class K_i. The average estimation of compactness on the training sample is calculated using the formula

$$F(E_0,\rho)=\frac{\sum_{i=1}^{2} m_i\theta_i}{m} \qquad (15)$$

where m - is the number of sample objects.

IV. Computational Experiment

The computational experiment was conducted with data of Echocardiogram [15]. 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. So 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).

The methods of visualization were considered over the three features spaces.

- original (11 features);
- expanded (14 features);
- reduced (10 features).

An expanded space from the original different in that each gradation of nominal feature considered (by the values {0,1}) as a separate quantitative feature. In the reduced space all nominal features (in our case the two nominal features) are presented a quantitative one using the linear mapping according to (7). The dimension of the space of the latent feature, which obtained by the nonlinear method, taking into account the above-mentioned three ways of describing the objects given in table 1.

Table 1.The dimension of space of the latent features

№	The space	The dimension (Values of weights of the features by (2))
1	Original	3 (0.748106, 0.452188, 0.324245)
2	Expanded	4 (0.787037, 0.517507, 0.382487, 0.335042)
3	Reduced	3 (0.748106, 0.452188, 0.324081)

The visual presentation of objects, obtained by linear [10] and nonlinear [11] methods in original feature space is given in figure 1 and figure 2. As a result of the nonlinear method (see table 1) in presentation used the first two latent features. The objects of the class K_1 marked with “x” symbol, K_2 - “o”.

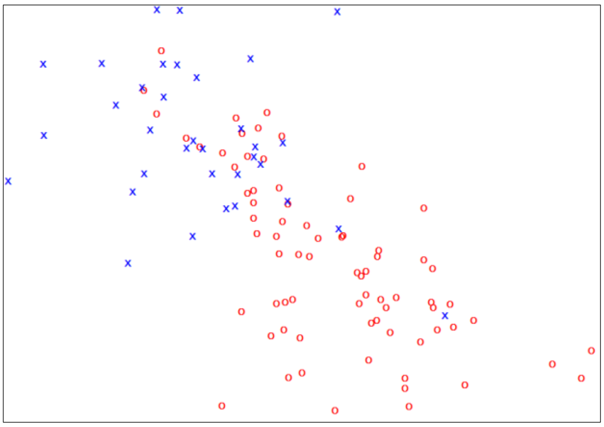

Fig.1. The visualization objects by linear method

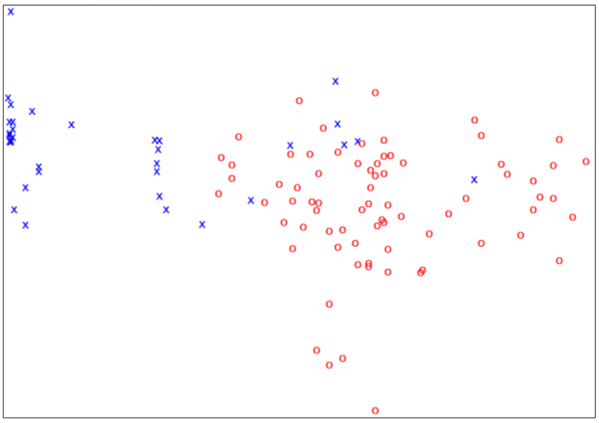

Fig.2. Visualization of objects by nonlinear method

From the image in figures, it is clear that objects of classes K_1 and K_2 better divided among themselves by using the nonlinear (Fig. 2) than the linear (Fig. 1) method.

Finding the groups is implemented through calculation the connectivity of objects. To demonstrate the connectivity of any two objects used a line which connects the two objects. The number of lines going out from an object means how many objects are connected to this object. And a group is the set of connected objects.

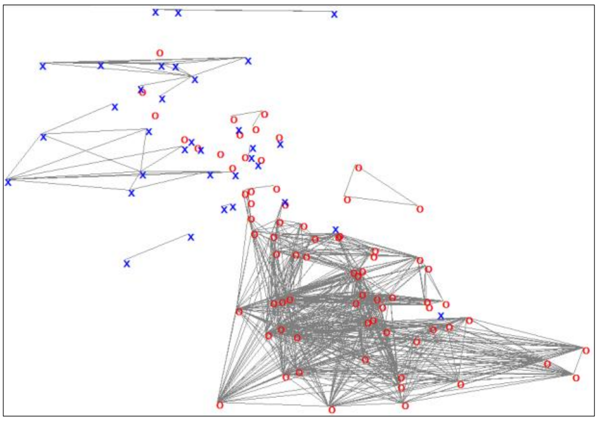

Fig.3. Grouping the objects according to the results of visualization by linear method

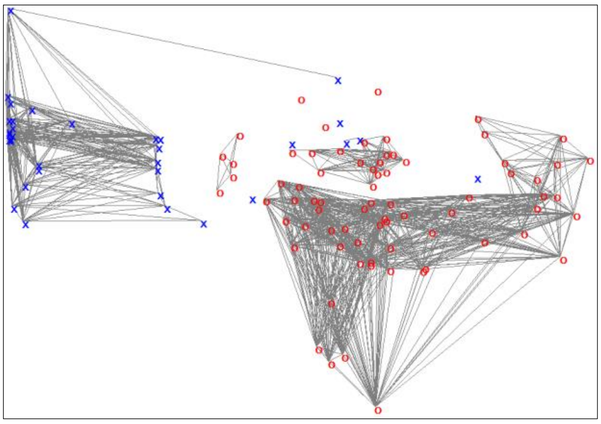

Fig.4. Grouping the objects according to the results of visualization by nonlinear method

As shown in figures 3, 4 the number of groups and its structure strongly influences when calculating the compactness. It is clear from the figure 4, that each class has equal number of groups, but the compactness are different: $F(K_1,\rho)=0.800949$ and $F(K_2,\rho)=0.636678$ (see table 5).

For comparative analysis, we consider the compactness of classes (14) and sample in whole (15) from the table 2, and mapping the objects by the nonlinear method in table 3. In brackets indicates the number of groups of objects and the dimension (see table 1) of the space of latent features.

Table 2. The compactness of the objects of classes (14) and sample as a whole (15)

The space	The compactness (number of groups) by		
	K_1	K_2	E_0
Original	0.491234 (3)	0.833910 (4)	0.599113 (7)
Expanded	0.448867 (4)	0.833910 (4)	0.570084 (8)
Reduced	0.921110 (4)	0.782006 (5)	0.877318 (9)

Table 3. The compactness of the objects of classes (14) and sample as a whole (15) by using the results of nonlinear method

The space (the number of latent features)	The compactness (number of groups) by		
	K_1	K_2	E_0
Original (3)	0.973338 (2)	0.737024 (4)	0.898943 (6)
Expanded (4)	0.800219 (4)	0.304498 (14)	0.644158 (18)
Reduced (3)	0.947041 (3)	0.782006 (5)	0.895086 (8)

It is clear from the table 2 and 3 that relatively low value of compactness is obtained when using values of features of the expanded space as a data.

The compactness indexes on the results of visualization of linear and nonlinear methods are given accordingly in table 4 and table 5. It have been used the first two latent features during the calculation by the results of nonlinear methods.

Table 4.The compactness by results of visualization using linear method

The space	The compactness (number of the groups) by		
	K_1	K_2	E_0
Original	0.598612 (14)	0.252595 (13)	0.489680 (27)
Expanded	0.323228 (11)	0.233564 (13)	0.295000 (24)
Reduced	0.686267 (8)	0.297577 (7)	0.563902 (15)

Table 5. The compactness by results of visualization using nonlinear method

The space	The compactness (number of the groups) by		
	K_1	K_2	E_0
Original	0.800949 (6)	0.636678 (6)	0.749234 (12)
Expanded	0.133674 (17)	0.429065 (13)	0.226667 (30)
Reduced	0.846603 (5)	0.647058 (5)	0.783783 (10)

To visualize by the linear method used mapping the objects by values of all features of the original space into two equivalent scales [10]. And the indexes of compactness (see table 4) were worse than using the nonlinear method (see table 5). The confirmation of this fact may be seen when comparing the images in figure 1 and figure 2.

To compare the results of linear and nonlinear methods with one of the well-known method as PCA, we used SPSS Statistics (Statistical Package for the Social Sciences) software package. There is the result of PCA method adequately consideration all factors (which eigenvalues greater than 1 [14]) and the first two factors in table 6.

Table 6. The compactness of the objects of classes (14) and sample as a whole (15) using PCA

The space	The compactness (number of groups) by		
	K_1	K_2	E_0
By all components	0.448502 (12)	0.243944 (11)	0.384104 (23)
By the first two components	0.600073 (8)	0.411764 (9)	0.540790 (17)

From the tables 5 and 6, we can see that using nonlinear method gives much better compactness by classes so and by sample as a whole than PCA method (see first row of the table 5 and the second row of the table 6).

The visual presentation of objects by PCA method using the first two factor (based on the first two eigenvalues) as a latent feature is presented in figure 5 and groups of objects are in figure 6.

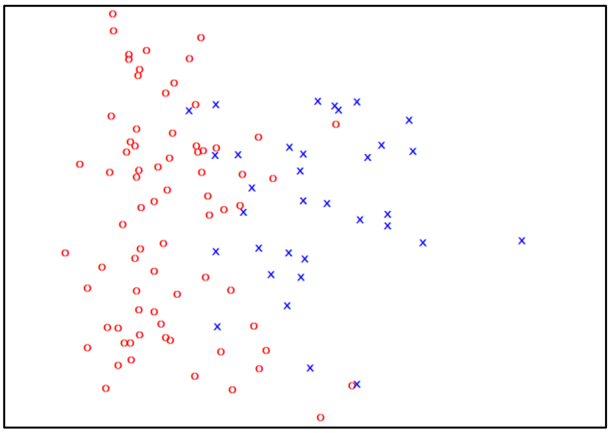

Fig.5. Visualization of objects by PCA

From the figure 5, we can see that when we use PCA method for visualization a lot of objects of different classes are intersect each other than using nonlinear method (see figure 2)

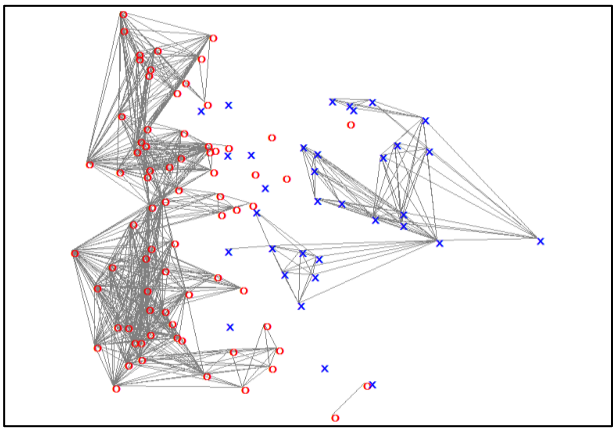

Fig.6. Grouping the objects according to the results of visualization by PCA

V. Conclusions

It is proved the effect of using the nonlinear method to form the space of latent features in the description of objects in pattern recognition problems. The indexes of effect serve an estimation of objects of the classes and the sample as a whole. The results of visualization may be used when searching the hidden regularities in databases.

References

[1] Duke V.A., *Methodology of finding the logical patterns in the domain of fuzzy systemology*: On the example of clinical and experimental studies, St. Petersburg, 2005.

[2] R. Duda, Hart P., Detection and Scene Analysis, Moscow: Mir, 1976.

[3] Zagoruyko N. G., Kutnenko O. A., Ziryanov A. O., Levanov D. A. *Obuchenie raspoznavaniyu obrazov bez pereobucheniya* // Mashinnoe obuchenie i analiz dannix. – 2014 . – T. 1 7. – pp. 891-901.

[4] Ignatev N.A. *Obobshennie otsenki i lokalnie metriki ob'ektov v intellektualnom analize dannix*. – Tashkent: Natsionalniy universitet Uzbekistana im. Mirzo Ulugbeka, 2014.

[5] H. Liu and H. Motoda, Feature selection for knowledge discovery and data mining, vol. 454, Springer Science & Business Media, 2012.

[6] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," Pattern recognition, vol. 43, no. 1, pp. 5-13, 2010.

[7] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 143-159, 2002.

[8] D. Asir Antony Gnana Singh, E. Jebamalar Leavline, R. Priyanka, P. Padma Priya, Dimensionality Reduction using Genetic Algorithm for Improving Accuracy in Medical Diagnosis, IJISA, vol. 8, №1, 2016, pp. 67-73.

[9] S. Dash,B. Patra, B.K. Tripathy, A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set, IJIEEB vol. 4, №2, 2012, pp. 43-50.

[10] Ignatiev N.A., On the construction of the feature space to search for logical regularities in pattern recognition problems, Computational technologies, vol. 17, №4, 2012, pp. 56-62.

[11] Ignatev N.A. Vichislenie obobshennix otsenok i ierarxicheskaya gruppirovka priznakov. Vestnik Tomskogo gosudarstvennogo universiteta. Tomsk, 2015, pp. 31-38.

[12] Saidov D.Yu., Nonlinear conversion of feature space and its analytical representation, International Youth Scientific Forum "Lomonosov-2015", 2015.

[13] Ignatiev N.A., Cluster analysis and selection of objects standards in pattern recognition problems with the teacher, Computational technologies, vol. 20, № 6, 2015, pp. 34-43.

[14] G. R. Norman, D. L. Streiner, Biostatistics:The Bare Essentials, PMPH-USA, 2008.

[15] http://archive.ics.uci.edu/ml/machine-learning-databases/echocardiogram

Authors' Profiles

Saidov Doniyor Yusupovich was born in Khorezm, Uzbekistan in 1986. He obtained his MSc (2011) and BSc(Ed) (2008) from the National university of Uzbekistan named after Mirzo Ulugbek. He is currently pursuing Ph.D degree at Algorithms and programming technologies department of the faculty mathematical science, National University of Uzbekistan, Tashkent, Uzbekistan. He has published over 7 refereed journal and conference papers in the areas of data mining. His recent publication lists as follow: Nonlinear conversion of feature space and its analytical representation (XXII international conference. Students, graduate students and young scientists "LOMONOSOV", Russian Federation, 2015), Grouping the features by the criterion compactness of objects of classes (Actual Problems of Applied Mathematics, computer science and mechanics, an international conference, 12-15 September, Voronej, Russian Federation, 2016), Analytical representation of recognition operators to calculate the generalized estimation(India: International Journal of Innovative Science Engineering and Technology, 2016), Generalizing ability of recognition algorithms taking into account the non-linearity (Computer Science and Energy Problems, Uzbekistan, 2016), Stability of the objects of classes and grouping the features (Problem of computational and applied mathematics,

Uzbekistan, 2016)
E-mail: doniyor_2286@mail.ru, doniyorsaidov86@gmail.com

How to cite this paper: Saidov Doniyor Yusupovich,"Data Visualization and its Proof by Compactness Criterion of Objects of Classes", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.8, pp.51-58, 2017. DOI: 10.5815/ijisa.2017.08.06