Baltic J. Modern Computing, Vol. 1 (2013), No. 1-2, pp. 9–28

Multidimensional Data Visualization Based on the Exponential Correlation Function

Laura RINGIENĖ, Gintautas DZEMYDA

Vilnius University, Institute of Mathematics and Informatics, Akademijos str. 4, LT-08663 Vilnius, Lithuania

lauraringiene@gmail.com, gintautas.dzemyda@mii.vu.lt

Abstract. Multidimensional data are often difficult to understand for a human because of their high dimensionality. Multidimensional data visualization is one of the ways for data perception where multidimensional data must be transformed in a low-dimensional space and presented visually for human decision. As a result of transformation there appear new data features, the number of which is lower than that of the original data features. In this paper, we present and investigate the way of reduction of dimensionality using the exponential correlation function, taking into account that there are clusters in the analysed set of multidimensional data.

Keywords: exponential correlation function, clustering, multidimensional scaling, visualization

1 Introduction

Multidimensional data usually describe objects (people, equipment, plant, nature, etc.), which are characterized by numerical features $x_1, x_2, \ldots, x_n$. The number m of the objects, that comprise a specific set of analysed objects, is finite. A certain collection of feature values describes one particular object $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1, m}$, of the analysed set, here n is the number of features. The object X_i can be interpreted as point, and the values $x_{i1}, x_{i2}, \ldots, x_{in}$ of features $x_1, x_2, \ldots, x_n$, in this case, are components of the point X_i. The analysed multidimensional data set can be described as a matrix $\mathbf{X} = \{X_1, X_2, \ldots, X_m\} = \{x_{ij}, i = \overline{1, m}, j = \overline{1, n}\}$, the ith row of which is the point $X_i \in R^n$ of the n-dimensional Euclidean space (Dzemyda *et al.*, 2013).

High dimensional data are difficult to understand for a human due to their voluminous n: to determine the structure, interrelations and groups of objects, etc. For that reason, there are many methods proposed for multidimensional data visualization. The visualization concept is rather wide, but we explore the methods of multidimensional data visualization that help to determine or estimate the structure of a set of multidimensional data objects (similarities between object groups, objects-outliers, and so on). There are two main groups of methods for visualizing multidimensional data:

direct visualization methods and projection methods, also called as dimension reduction techniques. Projection methods transform (project) the multidimensional data set $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}$ from the space R^n to a low-dimensional space R^d, $(d < n)$, where the obtain projection $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_m\} = \{y_{ij}, i = \overline{1,m}, j = \overline{1,d}\}$ of the data set $\mathbf{X}$ can be observed visually as $d = 1, 2$ or 3. The principal component analysis is a well-known linear projection method, whereas the multidimensional scaling is often used for nonlinear projection of multidimensional data (Dzemyda *et al.*, 2013).

The aim of this paper is to create a method to reduce the number of features of multidimensional data using the exponential correlation function, taking into account that there are clusters of similar objects in multidimensional data.

The method contains: 1) clustering of multidimensional data points into a certain number of clusters k, 2) transformation of n-dimensional data into a k-dimensional space R^k, and 3) visualization of the obtained k-dimensional data, using the projection method. Any method of projection (linear or nonlinear) can be used. The method has been investigated experimentally.

2 Visualization using multidimensional scaling

Multidimensional scaling (MDS) refers to a group of methods that are widely used for dimensionality reduction and visualization of multidimensional data (Borg and Groenen, 2005). The MDS method produces the projection $Y_i = (y_{i1}, y_{i2}, \ldots, y_{id})$ of the point $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ to a low-dimensional space R^d, $(d < n)$ (usually R^2 or R^3). After projecting to a low-dimensional space, similar objects (points) are arranged closer to one another while different objects (points) are located away from one another (Dzemyda *et al.*, 2013).

Let us denote the pairwise proximity of the points X_i and X_j by $d(X_i, X_j)$, and the distance between the corresponding points Y_i and Y_j in a low-dimensional space by $d(Y_i, Y_j), i, j = \overline{1,m}$. The aim of the MDS is to find the distances $d(Y_i, Y_j)$ as close as possible to $d(X_i, X_j)$. To this end, some least-squares objective function is minimized. The simple least-squares objective function, used in a literature, is called a *raw Stress* function and can be written as:

$$E_{rawStress} = \sum_{i<j} w_{ij}(d(Y_i, Y_j) - d(X_i, X_j))^2, \tag{1}$$

where w_{ij} are non-negative weights (Borg and Groenen, 2005). The simplest case is as $w_{ij} = 1$.

The application of MDS, as the least-squares objective function is *raw Stress* and $w_{ij} = 1$, is presented in Fig. 1. Four sets of multidimensional data, presented in the section Data of experiments, are visualized to the R^2 space, i.e. $d = 2$. We do not present labels and units for both axes in the figure because we are interested in observing the interlocation of points on a plane only. In Fig. 1 we see that there is one clearly separate cluster in Iris data, and there is no clear bound between the other two clusters. It is known that there are 5 clusters in Randomly generated data, but we can see only four clearly. There are no clear dividing boundaries between the clusters in the last two multidimensional data sets.

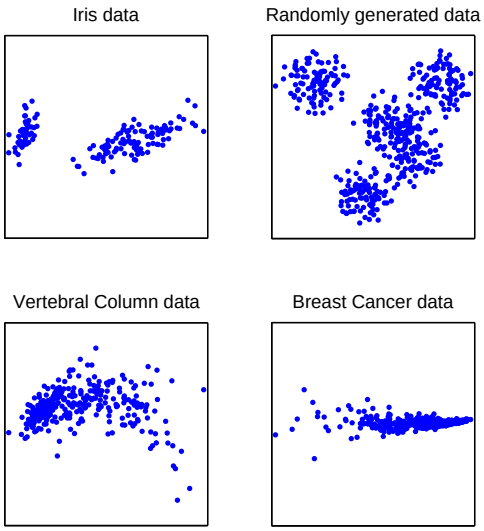

Fig. 1. Visualization of a multidimensional data sets by MDS as the least-squares objective function is *raw Stress*.

However the least-squares objective function presented in formula (1) is not the only one possible. There are more variants of this function. One of the examples is the *Stress*-1 function (Borg and Groenen, 2005), (Kruskal, 1964):

$$E_{Stress-1} = \sqrt{\frac{\sum_{i<j}(d(Y_i, Y_j) - d(X_i, X_j))^2}{\sum_{i<j}(d(Y_i, Y_j))^2}}, \tag{2}$$

The action of MDS, when the least-squares objective function is *Stress*-1, is presented in Fig. 2. If we compare the data of Fig. 1 with Fig. 2, we will note that visualization of Iris, Vertebral and Breast Cancer multidimensional data is unchanged, but there are 5 clear clusters in Randomly generated data.

MDS can be used directly for data visualization. However, there is an idea – maybe it is better first of all to perform a nonlinear transformation of multidimensional data by highlighting the clusters in the data, and afterwards to visualize the clustered data in order to see groups of objects better.

We give a short introduction of the clustering idea below.

One of the aims of visual data analysis is to find or even to see the clusters of data. In general, if we want to find clusters in the data and define their centers, we must use special methods meant for clustering. Clustering is the distribution of the analysed objects into different groups, also known as clusters, so that the objects in a group were similar to one another, and the objects in different groups were dissimilar (Dzemyda *et al.*, 2013). The data set $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}$ may be divided into non intersecting clusters $K_1, K_2, \ldots, K_k$ using any clustering method (k-means, classification tree, k nearest neighbour or others (Han *et al.*, 2011), (MacQueen, 1965), (Vesanto,

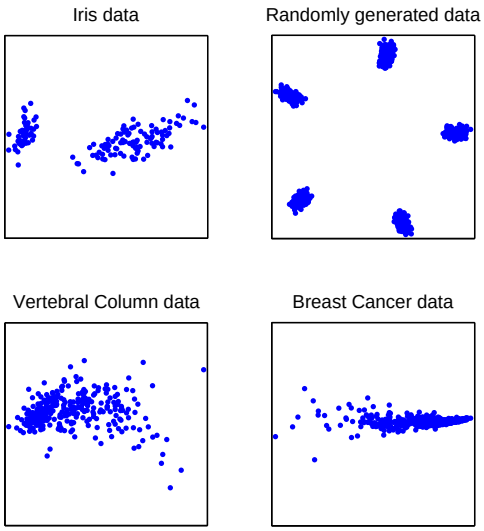

Fig. 2. Visualization of a multidimensional data sets by MDS as the least-squares objective function is *Stress*-1.

2001), (Dunham, 2003), (Cover and Hart, 1967)). In this paper, clustering is an inside procedure of the proposed method. We use here the k-means clustering method, which can find clusters $K_1, K_2, \ldots, K_k$ and the centers of clusters $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $\mu_j \in R^n$, $j = \overline{1,k}$ in the data set. With a view to achieve the objectivity of results, clustering has been carried out for several times in our experiments, because the function of clustering error is multiextremal and only the local, but not global minimum of the function is often found. The error of the k-means method is as follows:

$$E_k = \sum_{j=1}^{k} \sum_{X_i \in K_j} \| X_i - \mu_j \|^2, \tag{3}$$

where K_j is the jth cluster, $j = \overline{1,k}$, $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$ is the centre of the cluster K_j, $\mu_j \in R^n$, $\sum_{j=1}^{k} s_j = m$.

For illustration, the results of visualization of data consisting of matrix $\mathbf{X}$ and cluster centers $(\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $\mu_j \in R^n$, $j = \overline{1,k}$ using the MDS method, are presented in Figures 3 and 4. When comparing Fig. 1 with Fig. 3 and Fig. 2 with Fig. 4, we see that the location of points actually has not been changed, and the centers of the clusters μ_j, $j = \overline{1,k}$ (marked by $\circ$) are in the middle of clusters.

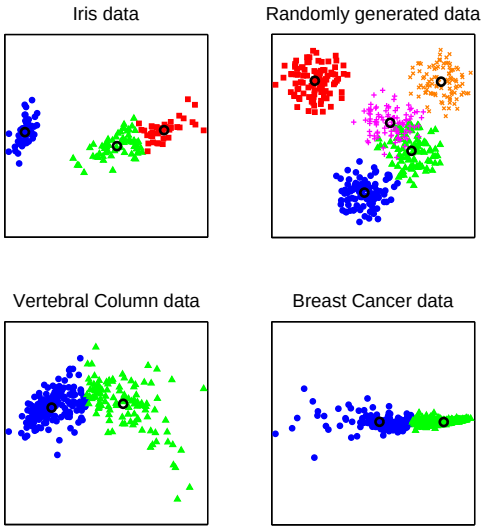

Fig. 3. Visualization of clustered multidimensional data by MDS using the *raw Stress* function.

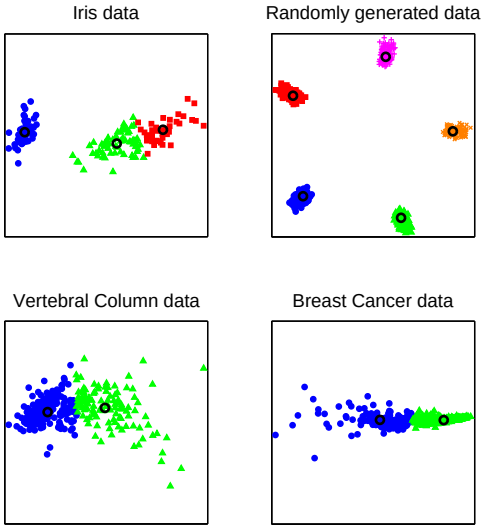

Fig. 4. Visualization of clustered multidimensional data by MDS using the *Stress*-1 function.

3 Application of the exponential correlation function to reduce the dimensionality of multidimensional data

The previous section presented the dimensionality reduction of multidimensional data into a low-dimensional space using the MDS method.

In this section, we present a method that includes:

- clustering of multidimensional data into a certain number k of clusters,
- transformation of n-dimensional data into the k-dimensional space R^k,
- visualization of k-dimensional data using nonlinear projection method (the MDS is used in this paper).

The advantage of the method is highlighting of the existing clusters in multidimensional data during visualization.

Let us discuss the proposed method in detail.

After clustering the data into a certain number of clusters k, we reduce the number of features n of multidimensional data $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1,m}$, where $X_i \in R^n$, by transforming $X_i \in R^n$ to $Z_i \in R^k : Z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$; here $k < n$. The dimensionality of $X = (x_1, x_2, \ldots, x_n)$ is reduced using some exponential correlation functions. We get a new data set $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_m\} = \{z_{ij}, i = \overline{1,m}, j = \overline{1,k}\}$, $k < n$, from the data set $\mathbf{X}$ using formulas:

A. Exponential correlation function (Yaglom, 1986):

$$z_j(X) = exp(-\gamma \parallel X - \mu_j \parallel), j = \overline{1,k}, \gamma = \frac{1}{2\sigma^2}, \tag{4}$$

B. Gaussian correlation function (Yaglom, 1986):

$$z_j(X) = exp(-\gamma \parallel X - \mu_j \parallel^2), j = \overline{1,k}, \gamma = \frac{1}{2\sigma^2}, \tag{5}$$

here μ_j is the center of the jth correlation function, $\mu_j \in R^n$, $\parallel X - \mu_j \parallel$ is the distance between the points X and μ_j, σ is the width parameter, which determines the function smoothness. Let us note that $\parallel X - \mu_j \parallel \geq 0$ and $\gamma > 0$. The function (5) is also called Gaussian radial basis function and is often applied in neural networks (Buhmann, 2003). The only difference between the exponential and Gaussian correlation functions is that in the Gaussian function the distance is squared.

After reducing the features from the number n to k, the obtained data set $\mathbf{Z}$ is visualized to the space R^2. Obviously, if the number of clusters $k > 2$, then we need to use projection methods to visualize the multidimensional data set $\mathbf{Z}$ to the space R^2. The MDS method is used in further experiments. In order to reveal the features of transformations (4) and (5) deeper, we visualized not only the data set $\mathbf{Z}$, but also the k-dimensional centers $\mu_j, j = \overline{1,k}$. Let us denote the obtained transformations of centers by $\mu_j^z = (\mu_{j1}^z, \mu_{j2}^z, \ldots, \mu_{jk}^z) \in R^k$. So, as in the previous section and in Figures 3 and 4, the total number of visualised points is $m + k$. The results are presented in Fig. 5. Two different functions of MDS have been used – *raw Stress* and *Stress*-1. The use of these two functions in the case of Randomly Generated data has given different results

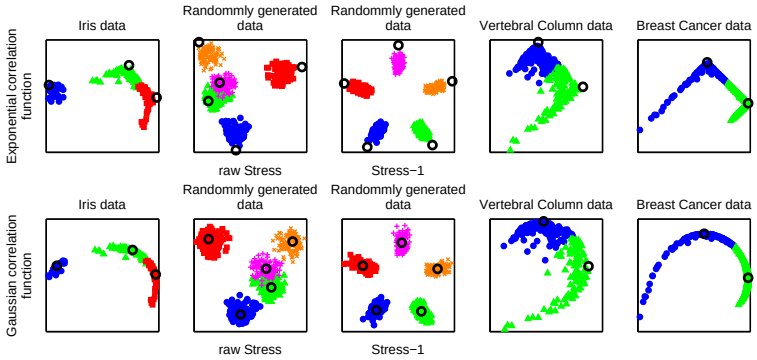

Fig. 5. Dimensionality reduction of multidimensional data sets using exponential and Gaussian correlation functions.

of visualization, but they did not affect the results of visualization of Iris data, Vertebral Column data and Breast Cancer data. The points corresponding to the points of different clusters are marked as • , ▲ , ■ , × , + . The cluster centers are marked by ○ .

Fig. 5 shows that we get different visualization results when we use different functions for transformation (4) or (5). There are two differences in visualization results when exponential and Gaussian correlation functions are applied: 1) Visualizing nature: visualization results are more angular in the case of exponential function, and they are more sleek in case of Gaussian function. 2) Location of the clusters center: in the case of Gaussian function, the centers are in the middle of the clusters, but in the case of exponential function the centers are shifted to the side from the points of the respective cluster and they acquires an exclusive property to be points where changing characteristic of cluster objects. This is due to that the exponential correlation function was analysed more in detail in this paper.

Fig. 5 shows that the points, visualized after dimensionality reduction using the exponential correlation function are located in two ways:

a) Isolated cluster (see visualization of Iris data and Randomly generated data). Points of the cluster make up a separate group. For example, a separate cluster in Iris data is marked by • . The points of this cluster focus in a clearly visible separate cluster. Isolated clusters in Randomly Generated data are also very clearly evident.
b) Close to each other clusters (see visualization of Iris data, Vertebral Column data and Breast Cancer data). Visualised points of a separate cluster scatter in the environment of two lines, which join together near the cluster center. The Breast Cancer data reflect best the arrangement of the objects in the environment of two lines (Fig. 5). This also is observed in the visualization of the Iris data and Vertebral Column data. The points that have similarities to that of the neighbouring cluster are visualized near the line that connects the centers of these neighbouring clusters.

When transforming multidimensional data from $X_i \in R^n$ into $Z_i \in R^k$, $i = \overline{1,m}$, using the exponential correlation function, it is important to choose the proper param-

eters of the function: centers μ_j and the width parameter σ. Just like most of the authors (Pierrefeu *et al.*, 2006), (Chang *et al.*, 2005), (Benoudjit and Verleysen, 2003), we choose the centers, by clustering data using the k-means method (Han *et al.*, 2011), (Vesanto, 2001), (MacQueen, 1965). The dependence of visualization results on σ is shown on Iris data as the number of clusters $k = 3$. The visualization dimensionality d is chosen equal to 3 and 2. Therefore, as $d = 3$, Z_i, $i = \overline{1,m}$, may be visualized directly, because $k = 3$. As $d = 2$, MDS was used to transform Z_i, $i = \overline{1,m}$, from R^3 to R^2. The results with different width parameter σ values (a) $\sigma = 0.3$; b) $\sigma = 3$; c) $\sigma = 30$; d) $\sigma = 100$) are presented in Fig. 6.

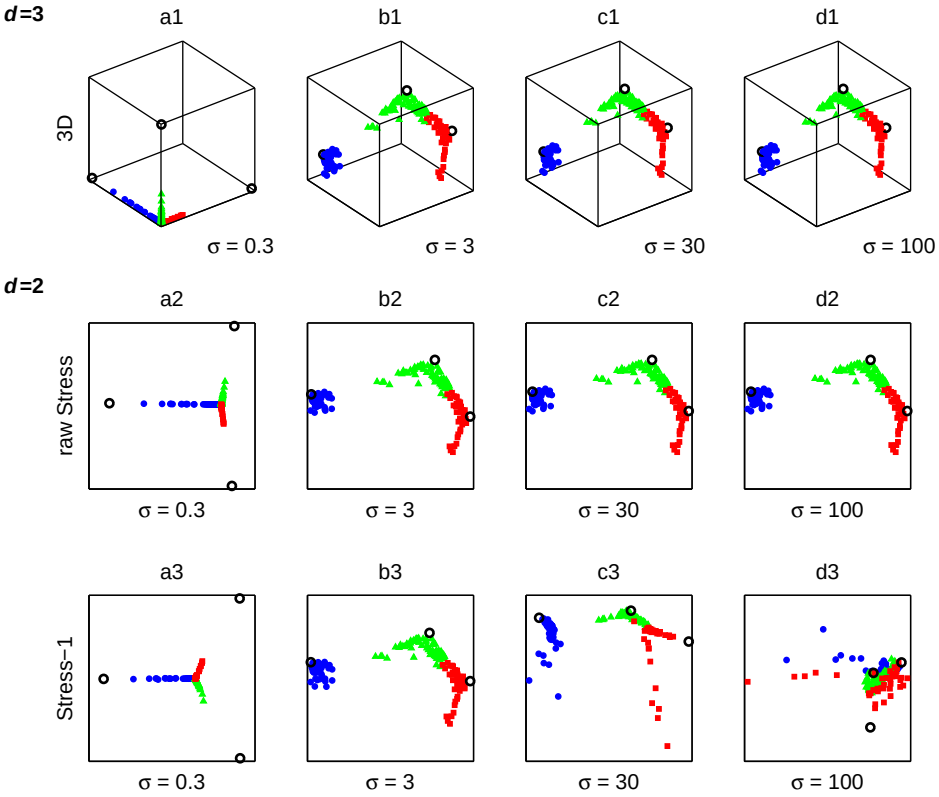

Fig. 6. Visualization results of Iris data with various σ using *raw Stress* and *Stress*-1

Looking over all the visualization results in Fig. 6 we can state that, if the width parameter of the correlation function is chosen too small (Fig. 6 (a1, a2, a3)), then all the points of the clusters are pushed to a totality and the centers of the clusters are outside of the clusters. If the width parameter of the exponential correlation function is selected properly (Fig. 6 (b1, b2, b3, c1, c2, d1, d2)), then the clusters are well distinguished. However, by visualizing the results into a two-dimensional space using MDS, as the

least-squares objective function is *Stress*-1, we notice that the width parameter can be choose too large (Fig. 6 (d3)), then the visualised data clusters overlap each other. We can conclude from Fig. 6 that, as $d = 3$ and $d = 2$ and MDS uses the *raw Stress* function, the width parameter σ does not have a significant influence on the relative location of visualised points, because the obtained Fig. 6 (b1, b2, c1, c2, d1, d2)) with different σ values visually look very similar. Only the scale of figures is different. However, when the *Stress*-1 is used (Fig. 6 (b3, c3, d3)), the width parameter σ from a certain range can become too large and it is very important to choose it properly.

On the other hand, it is interesting to observe the evolution of spread of the points on the two-dimensional plane with an increase of σ, when the *Stress*-1 is used. In this case, $\sigma = 0.3; 3; 30$. At first ($\sigma = 0.3$), the points concentrate in one large group, the cluster centers are on the sides and we see that they attract the points of their clusters. In this case, the belonging of the points to the clusters is determined by the location of the points between the cluster center and the center of the observed groups. The cluster which has the least similarity to the other clusters (points of this cluster are marked as •) separates from them when the width parameter σ increases. The other two clusters try to "pick out" the points typical of them only. Although in Fig. 6 (b3) we observe a continuous transition from one cluster to another, each of these clusters has separate points, which are located in another direction than towards the neighboring cluster. By increasing the width parameter σ up to 30 (Fig. 6 (c3)), we note that the attraction of the center of the middle cluster (its points are marked as ▲) becomes stronger and the points of the other two clusters start to move towards it. Here we see an increase in scattering of the points of the first cluster (its points are marked as •). The points of the third cluster (its points are marked as ▪) lose the contact with the center of their cluster and are shifted to the center of the second cluster (its points are marked as ▲). The points of the cluster, marked by ▲ , move to the center of their cluster.

All conclusion above are suitable and for Breast Cancer data. The width parameter σ values are different only a) $\sigma = 10$; b) $\sigma = 100$; c) $\sigma = 1000$; d) $\sigma = 10000$. The visualization results of Breast Cancer data with different width parameter σ values are presented in Fig. 7.

In further experiments, we use *Stress*-1 exclusively in MDS, because it is the only way when we can control the visualization results using various values of σ (see Fig. 6 and Fig. 7).

Fig. 6 and Fig. 7 illustrates that it is very important to choose the proper width parameter σ, but there is no single way to do that. However, we can acquire some experience from the radial basis function neural networks how to choose the proper width parameter σ. Neural networks of this type are widely applied in image recognition, classification, prediction, and solution of other problems.

The simplest approach is to take fixed radial-basis functions that define activation functions of the hidden units. The location of centers may be chosen randomly from the data set. This is considered to be a sensible approach, provided that the training data are distributed in a representative manner for the problem at hand (Lowe, 1989). However, it seems intuitively better, if the centers are chosen with regard to the clusters in multidimensional data. For the radial-basis functions, S. Haykin suggests to employ an isotropic Gaussian function whose standard deviation (i.e. width parameter σ) is

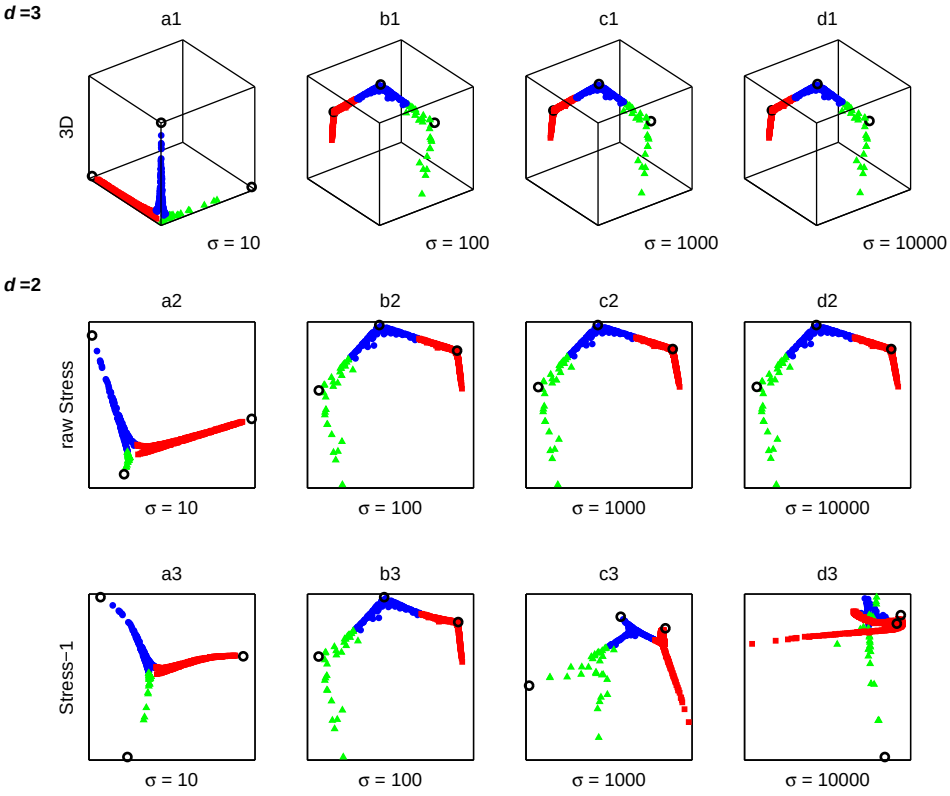

Fig. 7. Visualization results of Breast Cancer data with various σ using *raw Stress* and *Stress*-1

fixed according to the spread of the centers (Haykin, 2008). Specifically, a (normalized) radial-basis function centered at μ_j is defined as:

$$z_j(X) = exp\left(-\frac{\parallel X-\mu_j \parallel^2}{2\sigma_A^2}\right) = exp\left(-\frac{k}{d_{max}^2} \parallel X-\mu_j \parallel^2\right), j = \overline{1,k}, \quad (6)$$

where k is the number of clusters and d_{max} is the maximum distance between the chosen centers of all the clusters k. In effect, the width of all the Gaussian radial-basis functions is fixed as follows:

$$\sigma_A = \frac{d_{max}}{\sqrt{2k}} = \alpha d_{max}, \text{ where } \alpha = \frac{1}{\sqrt{2k}}. \quad (7)$$

This formula ensures that the individual radial-basis functions are not too peaked or too flat; both of these two extreme conditions should be avoided. As an alternative to formula (7) S. Haykin offers an idea to use individually scaled centers with broader widths in the areas of a lower data density, which requires experimentation with the training data (Haykin, 2008). But it is not specified how to do that.

This is leads us to use the exponential correlation function:

$$z_j(X) = exp\left(-\frac{\parallel X - \mu_j \parallel}{2\sigma_A^2}\right) = exp\left(-\frac{k}{d_{max}^2} \parallel X - \mu_j \parallel\right), j = \overline{1,k}. \quad (8)$$

Iris and Breast Cancer data, transformed into a low-dimensional space, where σ is calculated by formula (7) are presented in Fig. 8. On the basis of the results in Fig. 6 and Fig. 7, we can draw a conclusion, that the width parameter σ_A for the Iris data is proper, but the width parameter σ_A is too large for the Breast Cancer data, because we observe the movement of points of the first cluster (its points are marked by ▪) and the second cluster (its points are marked by •) towards the center of the third cluster (its points are marked by ▲). So the width parameter σ_A obtained by formula (7) is suitable not for all data.

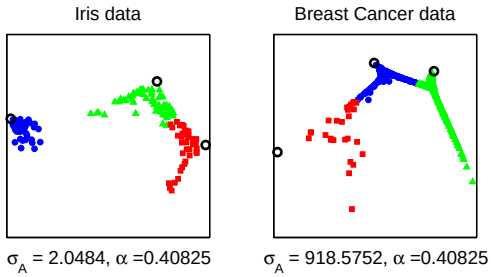

Fig. 8. Transformation results obtained calculating σ by formula (7).

For automatic selection of the width parameter, S. Haykin (2008) uses the maximum distance between the centers of clusters k. An alternative is the average distance between them. L. Pierrefeu *et al.* (2006) tests show that the width parameter σ calculated by the average distance between the centers of clusters gives some good results and seems to be well adapted. In fact, the average distance is not the optimal parameter for the width parameter. The best result is achieved with the width parameter approximately 20% smaller than the average distance. The method of choosing σ is quite simple:

1. Calculate the average distance between the centers of clusters:

$$d_{avg} = \frac{\sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} \parallel \mu_i - \mu_j \parallel}{k(k-1)}, \quad (9)$$

   where $\parallel \mu_i - \mu_j \parallel$ is the Euclidean distance between the centers μ_i and μ_j of clusters K_i and K_j, k is the number of clusters.
2. For the function

$$z_j(X) = exp\left(-\frac{\parallel X - \mu_j \parallel}{2\sigma_B^2}\right), \quad (10)$$

the width parameter is calculated as follows:

$$\sigma_B = \alpha d_{avg}, \text{ where } \alpha = \frac{1}{\beta}. \tag{11}$$

The authors (Pierrefeu *et al.*, 2006) propose to seek for the best value of σ_B by changing the value of β from 3.6 to 0.05 ($\alpha \in [0.28, 20]$) with a decrement of 0.05.

The results of dimensionality reduction, when σ_B is calculated by formula (11) using different α values (α changes from 0.28 to 20 with an increment of 3.944), are presented in Fig. 9 and Fig. 10. Based on Fig. 6 (a3, b3, c3, d3), we can state that the width parameter σ_B for the Iris data is a little bit too small (Fig. 9 (a)), because some movement of cluster points to cluster centers is noticed. The results of Fig. 9 (e, f) are very similar to that of Fig. 6 (d3), so we can state that the width parameter σ_B is too large because the visualized data clusters overlap each other, as $\alpha > 16$. The width parameter σ_B is also too large in Fig. 9 (c, d) because we observe the movement of the cluster points, marked by • and ▪, to the center of the cluster marked by ▲. If we reject the results when the width parameter σ_B is too small or too large for the Iris data, we can notice that the width parameter σ_B of the exponential correlation function is reasonable as $\alpha = 4.224$.

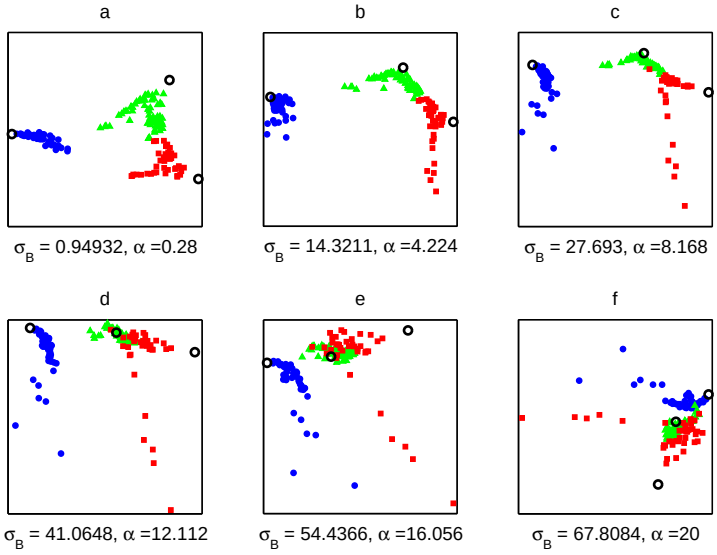

Fig. 9. The results of dimensionality reduction of Iris data using different α values.

Lets us compare the Breast Cancer data with the results of Fig. 7 (a3, b3, c3, d3). In the visualization results in Fig. 10 (b – f) we evidently see, that the visualised data clusters overlap each other. The experiments show that the best result has been got in Fig. 10 (a). The clusters are clearly distinguished in the data and the points of separate

clusters scatter in the environment of two lines. However, we observe that the visualised points of the cluster marked by ▪ lose the contact with projections of the center of their cluster.

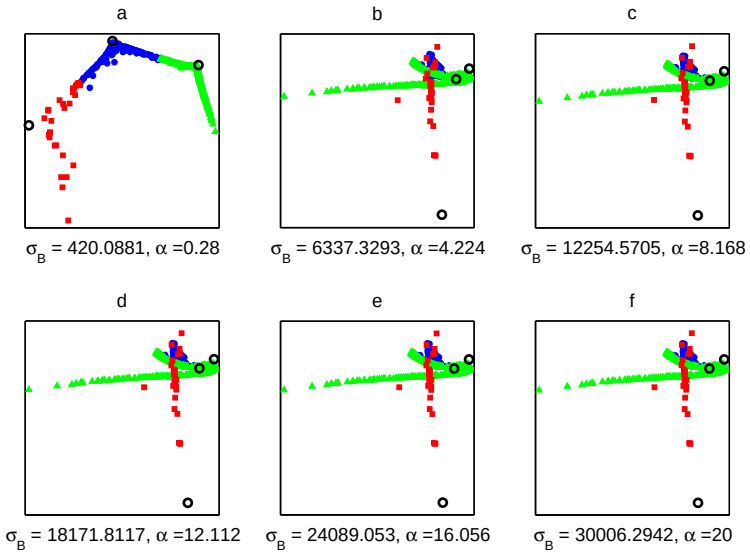

Fig. 10. The results of dimensionality reduction of Breast Cancer data using different α values.

There is rather a wide interval $[0.28, 20]$ to select α in formula (11), but it takes a lot of time to find a proper α from this interval if α runs with a small step. An example in Fig. 10 shows that the interval, proposed in (Pierrefeu *et al.*, 2006), is not suitable for the Breast Cancer data; it should be expanded. Thus, to find a proper α so that the width parameter σ_B were reasonable for transformation (10), we use the maximum distance τ from k minimal distances between the projections $\mu_j^y = (\mu_{j1}^y, \mu_{j2}^y)$, $j = \overline{1,k}$ of the centers $\mu_j^z = (\mu_{j1}^z, \mu_{j2}^z, \ldots, \mu_{jk}^z)$, $j = \overline{1,k}$ of clusters K_j, $j = \overline{1,k}$ and the projections $Y_i = (y_{i1}, y_{i2})$, $i = \overline{1,m}$ of the points $Z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$, $i = \overline{1,m}$.

$$\tau = max_{j=\overline{1,k}}\{min_{X_i \in K_j} \parallel Y_i - \mu_j^y \parallel\}. \tag{12}$$

The dependence of τ on α is presented in Fig. 11 for Iris dataset. At first with the increase of α, the value of τ decreases while α reaches some value α_b. Then the value of τ increases up to α reaches α_c, and afterwards it decreases again. For the sake of visualization, in Fig. 11, we present the results of visualization at the exceptional values of α ($\alpha = \alpha_a = 0.28, \alpha = \alpha_b = 2.28, \alpha = \alpha_c = 7.88$). Although the dependence of τ on α shows that there are smaller values of τ when the value of α is larger, but it is reasonable to fix the first found local minimum of τ, because $\alpha > \alpha_c$ is too large, as we see in Fig. 9 and Fig. 11.

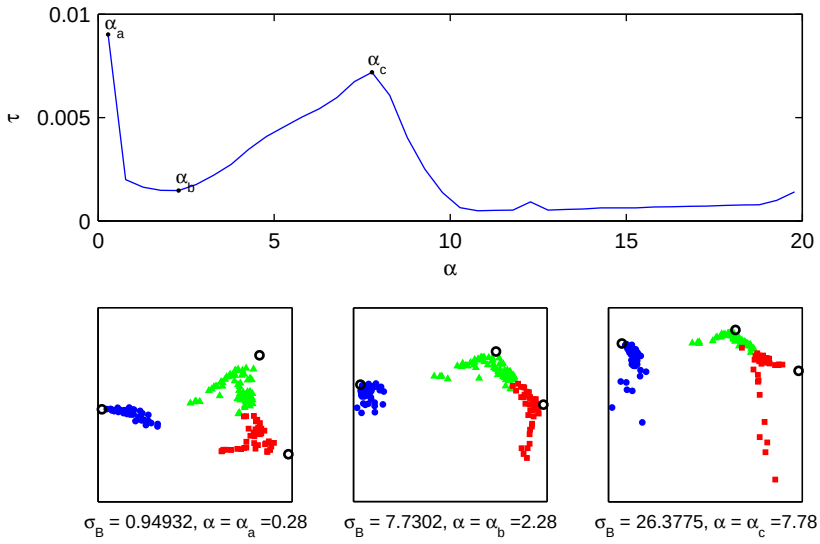

Fig. 11. Iris data: dependence of τ on α

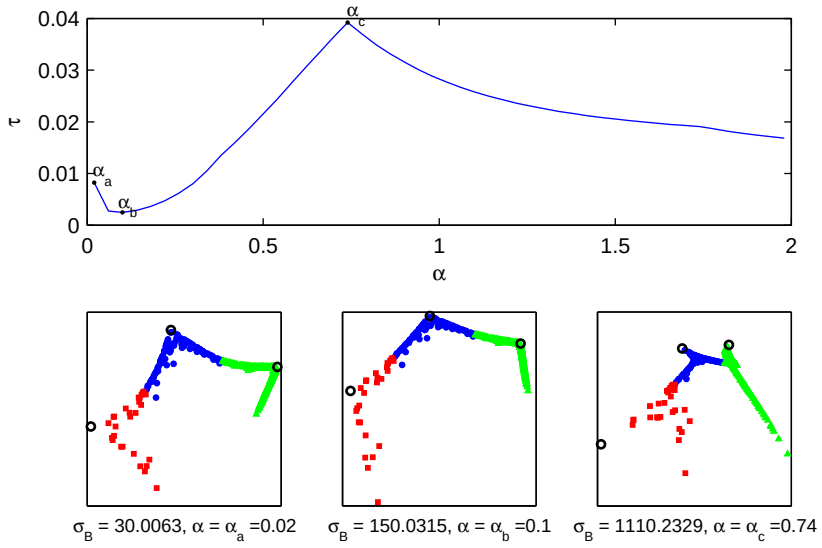

Fig. 12. Breast Cancer data: dependence of τ on α

As mentioned above, the interval $\alpha \in [0.28, 20]$ is not suitable for the Breast Cancer data, therefore we were looking for the parameter α in the interval $[0.02, 2]$. Dependence of τ on α and the results of visualization at the exceptional values of α ($\alpha = \alpha_a = 0.02, \alpha = \alpha_b = 0.1, \alpha = \alpha_c = 0.74$) are presented in Fig. 12.

The search for the minimal τ needs abundant calculations. To save the time for calculation, we tried to take the fixed α from formula (7), where α depends on the number of clusters, and the width parameter is calculated by formula (11):

$$\alpha = \frac{1}{\sqrt{2k}}, \sigma_C = \alpha d_{avg} = \frac{d_{avg}}{\sqrt{2k}}, \tag{13}$$

where d_{avg} is the average distance between the centers μ_j, $j = \overline{1,k}$ of clusters, k is the number of the clusters, $\alpha \in (0, 0.5]$, as $k \geq 2$.

The results, when σ_C is calculated by formula (13), are presented in Fig. 13. Comparing the transformation results of the Iris data in Figures 8, 9 and 13, we see that the width parameter σ_C is a bit too small, because attraction of the points to the projection centers μ_j^z of clusters is still going on.

σ_C for the Breast Cancer data is a little too large, because we observe some movement of the points of the first cluster marked by ▪ and the second cluster marked by • towards the center of the third cluster (its points are marked by ▲). We see that the results obtained by formula (13) are not the best ones, but, in this case, the view is better than in Fig. 8.

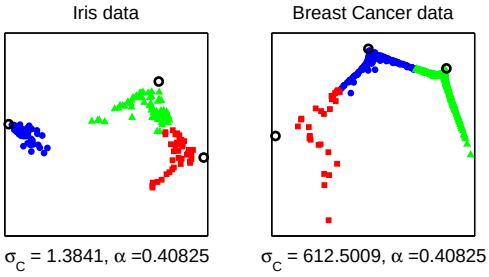

Fig. 13. Transformation results obtained calculating σ by formula (13).

4 Visual analysis of medical data

The dimensionality reduction of multidimensional data using the exponential correlation function, includes 1) clustering the multidimensional data points into a certain number of clusters k, 2) transformation of n-dimensional data into a k-dimensional space R^k, and 3) visualization of the obtained k-dimensional data, using the multidimensional scaling method. In this section, we present an experimental investigation of the method using the multidimensional medical data.

The method, discussed previously, was applied to the data grouped into the known number of clusters. However, the optimal number of clusters is usually unknown. Therefore the medical data (see the data sets 3 – 6 in the section Data of experiments) were divided into a different number of clusters. The direct visualization of multidimensional data, using the MDS method (matrix $\mathbf{X}$ and centers of the clusters $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $\mu_j \in R^n$, $j = \overline{1,k}$), is presented in the left column of Figures 14 – 17. The total number of the visualized points is $m + k$. Visualization of multidimensional data after transformation (matrix $\mathbf{Z}$ and centers of the clusters $\mu_j^z = (\mu_{j1}^z, \mu_{j2}^z, \ldots, \mu_{jk}^z)$, $\mu_j^z \in R^k$, $j = \overline{1,k}$), where the width parameter σ of the exponential correlation function was calculated by formula (11), is presented in the right column. The parameter α was calculated by formula (12). The total number of visualized points is $m+k$ as well. The points that correspond to different clusters are denoted by • , ▲ , ▪ , × . The projection of centers of the clusters are denoted by ○ .

Let us comment Figures 14 – 17. We see that after visualizing directly the multidimensional data set by the MDS method (Fig. 14 (a1, b1, c1); Fig. 15 (a1, b1, c1); Fig. 16 (a1, b1, c1); Fig. 17 (a1, b1, c1)), the location of the points in fact does not change, although the number of clusters changes. However, after the dimensionality reduction of multidimensional data using the exponential correlation function, the location of the points on the plane changes depending on the number of clusters (Fig. 14 (a2, b2, c2); Fig. 15 (a2, b2, c2); Fig. 16 (a2, b2, c2); Fig. 17 (a2, b2, c2)). Two-dimensional projections of the centers μ_j^z of clusters are shifted to the side from the points of the corresponding cluster, and the points of a separate cluster are sorted according to the similarity to the points of neighbouring clusters and to the inherent character typical only of a specific cluster.

The visualization results obtained using transformation (4) allow us to guess about the optimal number of clusters. For example, the Heart Diseases data (Fig. 16 (c2)) show that it is inappropriate to partition data into four clusters, because the points of the cluster marked by ▲ and the points of the cluster marked by × are similar to one another and do not distinguish the points typical of only of one cluster. The points of both clusters are visualised in one group, thus creating a single cluster. However, clustering into a higher number of clusters than optimal, some times may have advantages, because that allows us to look deeper into the data. For example, the Breast Cancer data. If we compare clustering into two (Fig. 15 (a1)) and four (Fig. 15 (c1)) clusters, we notice that four clusters are obtained by dividing the first two clusters (benign tumour and malignant tumour) into two parts (two clusters of benign tumoor and two clusters of malignant tumour). After transforming multidimensional data to the k-dimensional space, the points in the clusters are sorted in some way by similarity to the points of a neighbouring cluster. It enables us to interpret the data more accurately. It is very similar in the case of Vertebral Column data (Fig. 14). However, in the case of Parkinson Desease data (Fig. 17), when increasing the number of clusters only one cluster is divided (data about healthy people) into smaller clusters, ant the other cluster (data about sick people) (in Fig. 17 (a1) the points of the cluster are marked by ▲ ; in Fig. 17 (b1) by ▪ ; in Fig. 17 (c1) by ×) remains almost unchanged.

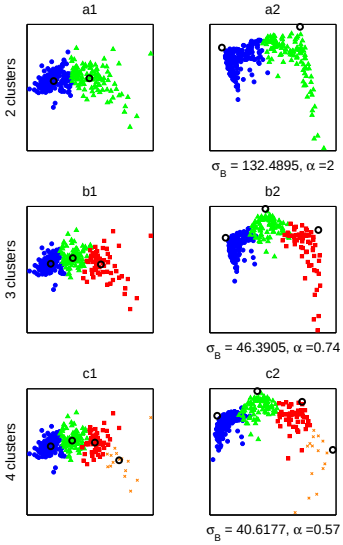

Fig. 14. Transformation results of Vertebral column with different number of clusters.

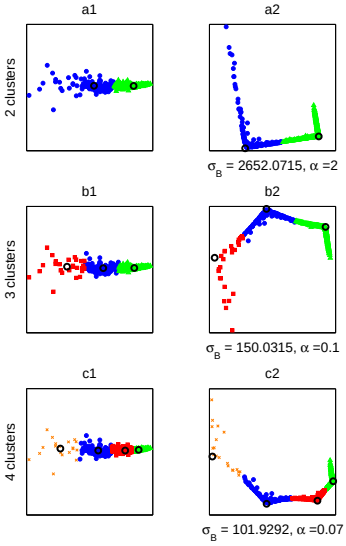

Fig. 15. Transformation results of Breast Cancer with different number of clusters.

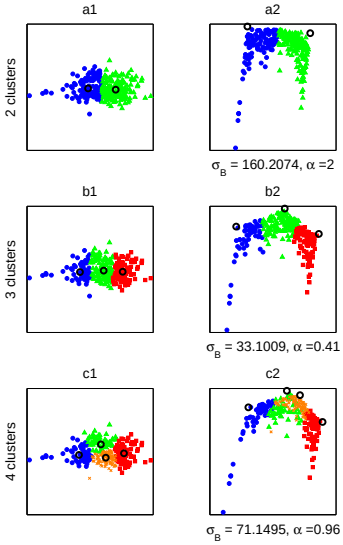

Fig. 16. Transformation results of Heart diseases with different number of clusters.

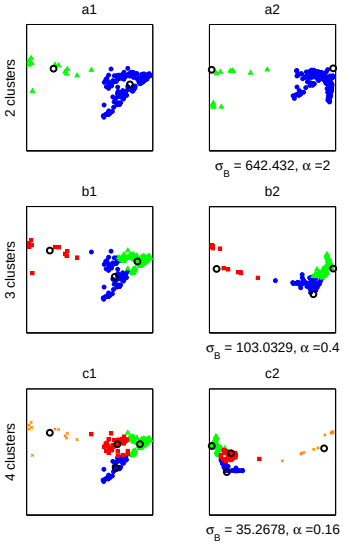

Fig. 17. Transformation results of Parkinson's diseases with different number of clusters.

If we compare a direct application of the MDS to the n-dimensional data with the results obtained after the transformation into a k-dimensional space and afterwards visualization by MDS, we can conclude that the proposed and investigated method allows us to predict better and evaluate the inherent character of the points of a particular cluster and similarities to other clusters.

5 Conclusions

Multidimensional data are often difficult to understand for a human because of their high dimensionality. Multidimensional data visualization is one of the methods for data perception where multidimensional data must be transformed into a low-dimensional space. As a result of transformation there appear new data features, the number of which is lower than that of the original data features. In this paper, we present and investigate the way of multidimensionality reduction using the exponential correlation function, taking into account that there are similar clusters in the analysed set of multidimensional data.

The method, proposed in this paper is based on the application of the exponential correlation function to dimensionality reduction of multidimensional data. The method includes:

- clustering of multidimensional data into a certain number k of clusters,
- transformation of n-dimensional data into the k-dimensional space R^k using the exponential correlation function,
- visualization of k-dimensional data using nonlinear projection method (the MDS was used in this paper).

The experimental investigation of the proposed method leads us to the following conclusions:

- The visualisation quality depends on:
  - the chosen number of clusters;
  - the chosen value of the width parameter σ;
  - the chosen least-squares objective functions in the multidimensional scaling method.
- Various ways of choice of the width parameter σ are discussed. The best parameter σ was defined applying formulas (11) and (12), but that requires the most computing time;
- The visualised points are located in two ways: isolated cluster (points of the cluster make up a separate group) and close to each other clusters (visualised points of a separate cluster scatter in the environment of two lines, which join together near the cluster center).

In fact, the method, proposed in this paper, is some extension of the nonlinear projection of multidimensional data, where data clustering and transformation of the clustered data to the lower dimensionality precedes the nonlinear projection. Such inclusion of clustering allows us to comprehend multidimensional data from a new standpoint.

6 Data of experiments

Six multidimensional data sets were used in the experiments. First five multidimensional data sets were taken from 'UCI Repository of Machine Learning Databases' (http://archive.ics.uci.edu/ml/).

1) *Iris Plants Database*. The data set consists of three kinds of flower iris – Setosa, Versicolour and Virginica ($k = 3$). There are 50 samples in each of the three classes, in total 150 ($m = 150$). Four features describe each flower of Iris – sepal length, sepal width, petal length and petal width ($n = 4$).
2) *Vertebral Column Database*. The data set containing values of six biomechanical features used to classify orthopaedic patients into 3 clusters ($k = 3$) – normal, disk hernia, or spondilolysthesis – or into 2 clusters ($k = 2$) – normal, abnormal. 310 patients comprise the whole data set ($m = 310$). Each patient is characterized by six biomechanical attributes: pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, and the grade of spondylolisthesis ($n = 6$).
3) *Breast Cancer Database*. The data set is grouped into 2 clusters ($k = 2$): malignant and benign. 569 tumours comprise the whole data set ($m = 569$). Each tumour is described by 30 features: radius (mean of distances from the center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area – 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" – 1) ($n = 30$).
4) *Heart Diseases Database*. The data set is divided into 2 clusters ($k = 2$) – the absence or presence of heart disease. 270 patients comprise the whole data set ($m = 270$). Each patient is characterized by 13 attributes: age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar $>$ 120mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, oldpeak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flouroscopy, thal (normal, fixed defect, reversable defect) ($n = 13$).
5) *Parkinson's Diseases Database*. The data set is classified into 2 clusters ($k = 2$) – absence or presence of Parkinson's disease. 195 patients comprise the whole data set ($m = 195$). 22 attributes characterize each patient that are typical of Parkinson's desease ($n = 22$).
6) *Randomly Generated Database*. Data are generated so that they make up 5 clusters ($k = 5$) of 100 points in each cluster, the whole data set consists of 500 points ($m = 500$) of R^{10} ($n = 10$). If the point belongs to a cluster, the numbered feature value is generated in the interval $[3, 5]$, the values of other features are generated in the interval $[-1, 1]$, i.e. $x_{ij} \in [-1, 1]$, and only if $X_i \in K_j$, then $x_{ij} \in [3, 5]$.

References

Benoudjit, N., Verleysen, M. (2003). On the Kernel Widths in Radial-Basis Function Networks. *Neural Processing Letters*, 18, 139–154.

Borg, I., Groenen, P. (2005). *Modern Multidimensional Scaling*, 2nd ed. Springer.

Buhmann, M. D. (2003). *Radial Basis Functions: Theory and Implementations.* Cambridge University Press.

Chang, Q., Chen, Q., Wang, X. (2005). Scaling Gaussian RBF kernel width to improve SVM classification. In: *International Conference on Neural Networks and Brain (ICNN&B '05)*, Vol. 1. IEEE Press, Beijing, pp. 19–22.

Cover, T. M., Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

Dunham, M. H. (2003) *Data Mining Introductory and Advanced Topics.* Pearson Education, Inc. Prentice Hall.

Dzemyda, G., Kurasova, O., Žilinskas J. (2013). *Multidimensional Data Visualization: Methods and Applications (Springer Optimization and Its Applications, Vol. 75)*. Springer.

Han, J., Kamber, M., Pei, J. (2011). *Data Mining: Concepts and Techniques (Morgan-Kaufmann Series of Data Management Systems)*, 3rd ed. Morgan Kaufmann.

Haykin, S. (2008). *Neural Networks and Learning Machines*, 3rd ed. Prentice Hall.

Kruskal, J. B. (1964). Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika*, 29 (1), 1–27.

Lowe, D. (1989). Adaptive Radial Basis Function Nonlinearities, And the Problem of Generalisation. In: *First IEE International Conference on Artificial Neural Networks (Conf. Publ. No. 313)*. London, pp. 171–175.

MacQueen, J. (1965). Some Methods for Classification and Analysis of Multivariate Observations. In: *Lecam, L. and Neyman, J. (Ed.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkley and Los Angeles, pp. 281–297.

Pierrefeu, L., Jay, J., Barat, C. (2006). Auto-adjustable method for Gaussian width optimization on RBF neural network. Application to face authentication on a mono-chip system. In: *The 32nd Annual Conference of the IEEE Industrial Electronics Society (IECON 2006)*. Paris, pp. 3481–3485.

Vesanto, J. (2001) Importance of Individual Variables in the k-means Algorithm. In: *Proceedings of PAKDD 2001*. Hong Kong, China, pp. 531–518.

Yaglom, A. M. 1986. *Correlation Theory of Stationary and Related Random Functions I: Basic Results (Springer Series in Statistics)*. New York: Springer.

Authors' information

L. Ringienė is a PhD student at the System Analysis Department of the Institute of Mathematics and Informatics, Vilnius University, Lithuania. She received her MSc in computer science from the Vilnius Pedagogical University, the Faculty of Mathematics and Informatics, Lithuania in 2008. Her research areas are visualization of multidimensional data and artificial neural networks.

G. Dzemyda, habil. dr., is a member of Lithuanian Academy of Sciences, professor, principal researcher and director of the Institute of Informatics and Mathematics of Vilnius University. His main research interests include optimization and visualization, data mining, and medical informatics.

Received February 26, 2013 , revised May 28, 2013, accepted June 5, 2013