VILNIUS UNIVERSITY

LAURA RINGIENĖ

HYBRID NEURAL NETWORK FOR MULTIDIMENSIONAL DATA VISUALIZATION

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07 T)

Vilnius, 2014

The doctoral dissertation was prepared at the Institute of Mathematics and Informatics of Vilnius University in 2008-2013.

Scientific Supervisor

Prof. Habil. Dr. Gintautas Dzemyda (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

The dissertation will be defended at the Council of the Scientific Field of Informatics Engineering at the Institute of Mathematics and Informatics of Vilnius University:

Chairman

Prof. Dr. Julius Žilinskas (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

Members:

Prof. Habil. Dr. Rimantas Barauskas (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07 T),
Prof. Dr. Vitalijus Denisovas (Klaipėda University, Technological Sciences, Informatics Engineering – 07 T),
Assoc. Prof. Dr. Olga Kurasova (Vilnius University, Physical Sciences, Informatics – 09 P),
Prof. Habil. Dr. Leonidas Sakalauskas (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

Opponents:

Prof. Dr. Eduardas Bareiša (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07 T),
Prof. Dr. Dalius Navakauskas (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07 T).

The dissertation will be defended at the public meeting of the Council of the Scientific Field of Informatics Engineering in the auditorium number 203 at the Institute of Mathematics and Informatics of Vilnius University, at 1 p. m. on 5$^{\text{th}}$ of September 2014.

Address: Akademijos st. 4, LT-08663 Vilnius, Lithuania.

The summary of the doctoral dissertation was distributed on the 4$^{\text{th}}$ of August 2014.

A copy of the doctoral dissertation is available for review at the Library of Vilnius University.

VILNIAUS UNIVERSITETAS

LAURA RINGIENĖ

HIBRIDINIS NEURONINIS TINKLAS DAUGIAMAČIAMS DUOMENIMS VIZUALIZUOTI

Daktaro disertacijos santrauka
Technologijos mokslai, informatikos inžinerija (07 T)

Vilnius, 2014

Disertacija rengta 2008-2013 metais Vilniaus universitero Matematikos ir informatikos institute.

Mokslinis vadovas

prof. habil. dr. Gintautas Dzemyda (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Disertacija ginama Vilniaus universiteto Matematikos ir informatikos instituto Informatikos inžinerijos mokslo krypties taryboje:

Pirmininkas

prof. dr. Julius Žilinskas (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Nariai:

prof. habil. dr. Rimantas Barauskas (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07 T),
prof. dr. Vitalijus Denisovas (Klaipėdos universitetas, technologijos mokslai, informatikos inžinerija – 07 T),
doc. dr. Olga Kurasova (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P),
prof. habil. dr. Leonidas Sakalauskas (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Oponentai:

prof. dr. Eduardas Bareiša (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07 T),
prof. dr. Dalius Navakauskas (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Disertacija bus ginama Vilniaus universiteto viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2014 m. rugsėjo mėn. 5 d. 13 val. Vilniaus universiteto Matematikos ir informatikos instituto 203 auditorijoje.

Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2014 m. rugpjūčio mėn. 4 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje.

1. Introduction

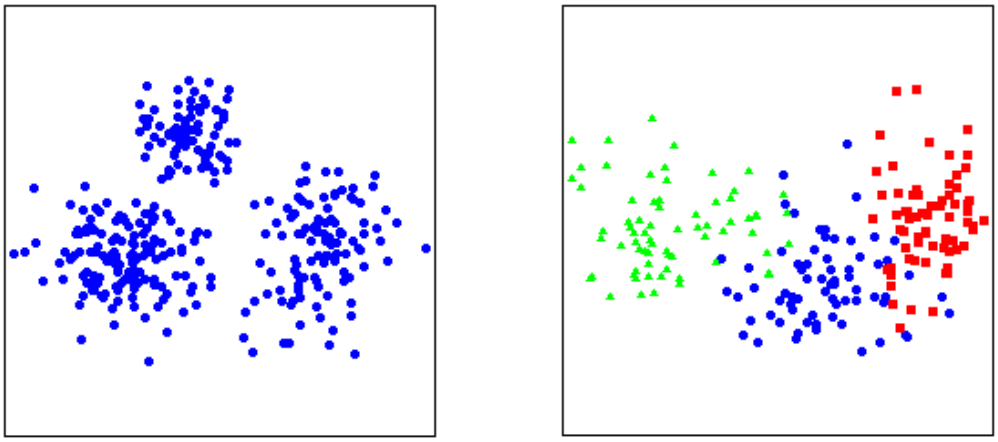
1.1. Research Area

The amount of stored data is growing with a rapid development of modern technologies in various areas: engineering, economics, medicine, ecology and many others. Data acquition was necessary for new knowledge. For example, to predict future activities, to identify the critical cases, and to generalize. However, it is complicated for a human to understand and interpret a very large amount of data. Therefore, various data mining methods are developed for solving important tasks: data partition into groups, definition of data structure, finding of relations or uniqueness, and so on. Multidimensional data visualization using two-dimensional or three-dimensional space helps to find a solution to the mentioned tasks. The area of research is data mining based on multidimensional data visual analysis.

1.2. Relevance of the Dissertation

In this dissertation, we consider such multidimensional data that describes the set of objects (people, equipment, products of manufacturing, natural phenomena, plants, etc.) which are characterized by the same features (parameters, attributes). A combination of values of all features characterizes a particular object $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1,m}$, where n is the number of features, m is the number of objects, and i is the order number of the object. Often, X_i are interpreted as n-dimensional points, and features $x_1, x_2, \ldots, x_n$ are coordinate values of points X_i. The analyzed data set can be described as matrix $\mathbf{X} = \{X_1, X_2, \ldots, X_m\} = \{x_{ij}, i = \overline{1,m}, j = \overline{1,n}\}$, where the ith row of this matrix is a point of Euclidian space $X_i \in \mathbb{R}^n$ (Dzemyda ir kt., 2013).

There are many methods for multidimensional data visualization, but in order to facilitate data interpretation and understanding, they are further developed rapidly (Dzemyda ir kt., 2013). Also, these methods are realized in many software systems: Matlab (R2009b, The MathWorks, http://www.mathworks.se/), Orange (Podpečan ir kt., 2012), Weka (Hall ir kt., 2009), and others. Methods of visualization provide the available multidimensional data to a human in an apprehensible space (two-dimensional or three-dimensional) preserving the point arrangement, i.e. keeping their similarities and differences. There is a need to visually assess the data set structure and properties: generated groups, impressively distinctive objects, object similarities/differences and so on. In most cases of data, the groups of objects are not clearly separated, i.e. a boundary between the groups of objects is not visible, as shown in Figure 1.1a, where the E.coli data set is presented on the plane using multidimensional scaling (Borg ir Groenen, 2005). The data set consists of 336 E.coli bacteria, described by 7 features. We see three groups of objects (bacteria), although there are much more groups, in fact. Usually different groups of objects are folded or even the objects of one group fall into another group of objects. For example, Wheat seeds data set is visualized on the plane in Figure 1.1b. The data set consists of 210 wheat seeds, described by 7 features. The objects of different groups are represented in different colours. There is a need to separate one group from another or discover groups of objects. For example, there can be a need to find wheat seeds in every group that have more similarities with the seeds of another group, or vice versa, to purify the seeds of a specific group.



(a) E.coli **(b)** Wheat seeds

Figure 1.1 Examples of multidimensional data visualization

After data visualization by the classical methods, if we add a new object to the data set, we have to solve the visualization problem again applying the visualization method to the whole data set including the new object. The specific methods to find a place of the new object among the visualized ones (for example, triangulation method (Karbauskaitė ir Dzemyda, 2006)) may be used, too. Artificial neural networks are successfully used for the arrangement of the new points on the plane.

1.3. The Objective and Tasks of the Research

The objective of the dissertation is to create a method for making a multidimensional data projection on the plane such that the researcher could see and assess the intergroup similarities and differences of multidimensional points.

In order to achieve the target, the following tasks are solved:

1. An analytical overview of mining methods of the data of the work-related objective: visualization methods, clustering methods, and artificial neural networks, as well as the developed compounds of the radial basis functions and multilayer perceptron;
2. Analysis of the artificial neural network possibilities to visualize multidimensional data;
3. Optimization of the radial basis function applicability to multidimensional data dimension reduction, based on the visual analysis of the obtained results;
4. Investigation of the compound of radial basis functions and a multilayer perceptron (hybrid neural network REGM) for visual exploration of multidimensional data, in order to assess intergroup similarities/differences;
5. Development of visualization quality criteria that would help to evaluate the obtained visualization results;
6. Development of criteria for selection of the well-trained network REGM.

1.4. Scientific Novelty

A new hybrid neural network is proposed and investigated in this dissertation. This neural network integrates the ideas both of the radial basis function neural network and that of a multilayer perceptron, which has the properties of a "bottleneck" neural network. Namely this is the scientific novelty of the dissertation. In the sequel this network will be called a network REGM. Further, we present detailed analysis of this idea.

The network REGM consists of two parts: the radial basic function layer and the multilayer perceptron of the special structure. The radial basis functions perform a certain transformation of points from the n-dimensional space $\mathbb{R}^n$ into the space $\mathbb{R}^k$ of the desired dimensionality k, $k < n$. Gaussian and exponential functions are considered. The width parameter is used to calculate the values of such radial basis functions in neural networks. In the literature, it is proposed to choose the width parameter by the network-related error. In the proposed network REGM, the width parameter is chosen before the training of the whole network. Therefore, we have to look for other ways to select the width parameter. In the dissertation, we suggest to choose the width parameter by the scattering of objects in the clusters and the average distance between their centers. Application of radial basis functions to data dimensionality reduction to k was optimized, based on the visual analysis of the obtained results.

A special structure multilayer perceptron is in the second constituent part of network REGM. The last hidden layer of this multilayer perceptron consists of a small number of neurons (2 or 3). The purpose of the network REGM is to project the multidimensional data into a two-dimensional or three-dimensional space (the projection is obtained in the last hidden layer), where the points corresponding to the objects can be monitored visually. The properties of clusters are also revealed in the visualized data. The knowledge about cluster composition and objects making up the clusters is got before training the network REGM and is used during the training.

After training the network REGM, the visually presented projections of multidimensional data are evaluated by the several visualization quality criteria in this dissertation. To achieve the better visual representation of data, it is reasonable to train the network REGM for several times and to choose the best projection. The selection criteria are proposed to compare the solutions and to find faster the projection confirming to the visualization quality criteria.

1.5. Defending Propositions

1. Consolidation of ideas of the radial basis function neural network and a special multilayer perceptron allows us to find the projection on the plane such that the researcher can see and evaluate intergroup similarities/differences of multidimensional points.

2. The width parameter of radial basis functions for the network REGM can be determined according to the scattering of objects in clusters and the average distance between centers of the clusters.

3. The proposed three visualization quality criteria evaluate the visualization results performance of the trained network REGM.

4. The proposed two selection criteria facilitate selection of the best projection of the data set, if the network REGM is trained for several times. Using them selection can be automated.

1.6. Approbation and Publications of the Research

The main results of the dissertation were published in 3 periodical scientific publications. The main results of the work have been presented and discussed at 5 national and international conferences.

1.7. Outline of the Dissertation

The dissertation is written in Lithuanian. The dissertation consists of 5 chapters and references. The chapters of the dissertation are as follows: Introduction, Data mining methods, related to the aim and tasks, Network REGM for multidimensional data visualization, Experimental research, Summary and general conclusions. The dissertation also includes the list of notation and abbreviations. The scope of the work is 130 pages that include 59 figures and 32 tables. The list of references consists of 101 sources.

2. Data Mining Methods Related to the Aim and Tasks

Visualization and clustering methods of multidimensional data are analytically overviewed in this section. Artificial neural networks are analyzed, which are used for multidimensional data visualization:

- Multilayer perceptron. Multidimensional data are visualized by the multidimensional scaling method, then a multilayer perceptron is trained using the results. As a result: the network is able to obtain a low-dimensional projections of new multidimensional points that were not visualized using MDS.
- The neural network of SAMANN type. This is a specific feed-forward neural network, which realizes Sammon's mapping in an unsupervised way. Low-dimensional projections of the data are obtained on the network output.
- A "bottleneck" type neural network. The idea of this network is as follows: what is presented into the network input is to be obtained on the output. The projection of data is obtained in the middle hidden layer which consists of two or three neurons.
- Self-Organizing Map. This network is trained in an unsupervised way. This network not only finds projection of the data set on the plane, but also divides the existing data into clusters.
- Radial basis function neural network. This network classifies data and searches for their projection on a hypercube in the hidden layer.

Operation strategies of the mentioned neural networks, except the RBF network, are oriented to saving distances between the points when looking for the projection of multidimensional data on the plane. Depending on the optimization criterion, distances between the closer of father points may be tried to be saved.

The analytical survey of hybrid neural networks (various radial basis functions and multilayer perceptron compounds) has showed that networks of this type are constructed in many different fields and for solving specific tasks: complicated data classification (diversely interlaced clusters, for example: spirally), creation of an equalizer, modelling of microwave devices, finding of spatial interpolation. The results of hybrid networks are more accurate in comparison with the results of radial basis function neural networks or multilayer perceptrons. The hybrid neural network architecture for solving a specific task is chosen according to the individual characteristics of separate networks.

The analytical review has showed the advantages and specificity of certain decisions:

1. The radial basis function neural network realized the possibility to evaluate clusters in the considered data, when the centers of certain individual clusters are used in different radial basis functions. Each radial basic function is "sensitive" to one particular cluster center.

2. The projection of multidimensional data is obtained in the hidden layer of neurons in the "bottleneck" neural network.

3. The knowledge of a particular point is used for training the multilayer perceptron in a supervised way.

The analysis of this section has showed that when looking for the data projection in which the researcher could discover intergroup similarities/differences of points, it is necessary to combine different types of neural network properties and not try to keep the distance between the points in the projection. The knowledge of data clusters is very important for training a network suggested in this dissertation. The knowledge can be obtained by clustering data using clustering methods.

3. The Network REGM for Multidimensional Data Visualization

This dissertation aims to develop a network that could be trained by presenting multidimensional data to be visualized to the trained network inputs. The specific reaction, associated with certain characteristics of the data, i.e. their belonging to clusters, is required on the output.

A model of combination of the new hybrid network (REGM) of radial basis functions and multilayer perceptron is presented in this section. Training peculiarities are discussed and visualization quality criteria are proposed.

3.1. Assumptions to Develop a New Visualization Method

As already mentioned in the previous section, the developed methods for multidimensional data visualization, try to keep the distance between the points when searching for data projection on the plane. This section presents an idea how to transform multidimensional data so that the intergroup point similarities in the projection were more noticeable.

There is an idea to reduce the number of features n of multidimensional data expressed by points $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1, m}$, where $X_i \in \mathbb{R}^n$, by transforming

$X_i \in \mathbb{R}^n$ to $Z_i \in \mathbb{R}^k$: $Z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$; where $k < n$. The dimensionality of point $X = (x_1, x_2, \ldots, x_n)$ is reduced using a certain radial basis function that is associated with a particular data cluster. We get a new k-dimensional data point $Z = (z_1, z_2, \ldots, z_k)$, $k < n$, using the following formulas:

1. The exponential function:

$$z_j(X) = \exp(-\gamma \parallel X - \mu_j \parallel), j = \overline{1,k}, \gamma = \frac{1}{2\sigma^2}, \tag{3.1}$$

2. The Gaussian function:

$$z_j(X) = \exp(-\gamma \parallel X - \mu_j \parallel^2), j = \overline{1,k}, \gamma = \frac{1}{2\sigma^2}, \tag{3.2}$$

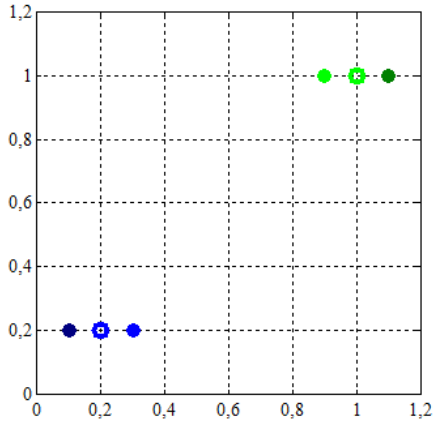
here μ_j is the center point of the jth function, $\mu_j \in R^n$, $\|X - \mu_j\|$ is the distance between the points X and μ_j, σ is the width parameter which determines the function smoothness. Note that $\|X - \mu_j\| > 0$ and $\gamma > 0$. The only difference between the exponential and Gaussian functions is that in the Gaussian function the distance is squared. The new data set $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_m\} = \{z_{ij}, i = \overline{1,m}, j = \overline{1,k}\}$ is obtained from the data set $\mathbf{X} = \{X_1, X_2, \ldots, X_m\} = \{x_{ij}, i = \overline{1,m}, j = \overline{1,n}\}$ based on formulas (3.1) or (3.2), i.e. nonlinear transformation of data set $\mathbf{X}$ is performed, where the clusters of this data set are taken into account (Ringienė and Dzemyda, 2013).

For the sake of simplicity, let us take $n = 2$ and $k = 2$. Let us consider the point position on the plane after the data transformation. Let us take the data set $\mathbf{X}$, which consists of 6 data points ($m = 6$) as an example. The data set is presented in Table 3.1, i.e. the initial data. The data distribution on the plane is shown in Figure 3.1a, which illustrates that the data set consists of two clear clusters. One cluster is marked in blue, the another cluster is green. The middle point of each cluster is the center of the cluster μ_j, marked as blue or green circles.

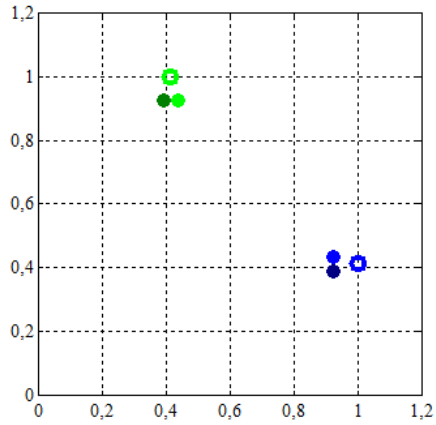
Table 3.1: The data set and the results after its transformation

No.	Data set $\mathbf{X}$		Exponential transformation		Gaussian transformation	
	x_1	x_2	z_1	z_2	z_1	z_2
1	0.1	0.2	0.90	0.29	0.99	0.23
2	0.2	0.2	1	0.32	1	0.27
3	0.3	0.2	0.90	0.34	0.99	0.32
4	0.9	1	0.34	0.90	0.32	0.99
5	1	1	0.32	1	0.27	1
6	1.1	1	0.29	0.90	0.23	0.99

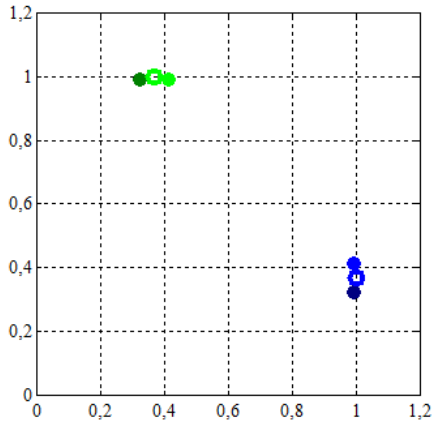
The results obtained after transformation by the exponential or Gaussian function are presented in Table 3.1. The results are presented visually in Figures 3.1b and 3.1c. Figure 3.1b shows that in the case of the exponential function, the cluster centers are separated from the other cluster points, and the remaining cluster points converge. The points, having a similarity with the neighbouring cluster points, appear closer to the neighbouring cluster. The points marked by lighter shades (blue and green) have more similarity to one another than that marked by darker shades.



(a) Initial data



(b) Transformation performed by the exponential function



(c) Transformation performed by the Gaussian function

Figure 3.1 The data set presented visually

The distribution of cluster points in the environment of the cluster center changes also in the case of the Gaussian function (Figure 3.1c). The change of the position of points, which are not the cluster centers, is similar to that in the case of the exponential function, but not so distinct.

3.2. A Model of the Network REGM

A new hybrid neural network is proposed and investigated in this dissertation. This neural network integrates the ideas of the radial basis function neural network and that of the multilayer perceptron, which has the properties of the "bottleneck" neural network. This network is called a network REGM.

The name of REGM is up of the first letters of its component neural networks and transformation function names (i.e. ***R****adial basis function neural network*, ***E****ksponential function*, ***G****aussian function*, ***M****ultilayer perceptron*).

The network consists of two parts that express separate training stages. The first part is a certain transformation of the points from the n-dimensional space $\mathbb{R}^n$ into the desired dimension space $\mathbb{R}^k$, $k < n$. A multilayer perceptron of special structure is in the second part. The last hidden layer of this multilayer perceptron consists of a small number of neurons (2 or 3). If there are more neurons in the output layer than in the last hidden layer, it resembles the "bottleneck" neural network in some

sense. However, this is only a very distant analogy, because there is a symmetry in the "bottleneck" neural network and one try to get, what was given to the network during the training, on the output.

The network REGM is used for the visual analysis of multidimensional data in such a way that the output values of the neurons of the last hidden layer are the two-dimensional or three-dimensional projections of the n-dimensional data, when the set $\mathbf{X}$ is filed to the network.

A peculiarity of the network is that the visualization results on the plane reflect the general structure of the data (clusters, proximity between clusters, intergroup similarities of points) rather than the location of multidimensional points. Note that multidimensional data clustering results can be used not only for calculating the radial basis function parameters, but also for the visual presentation of the results on the plane. Different clusters of points on the plane, painted in different colours, provide additional knowledge to the researcher. A scheme of the neural network REGM is presented in Figure 3.2.

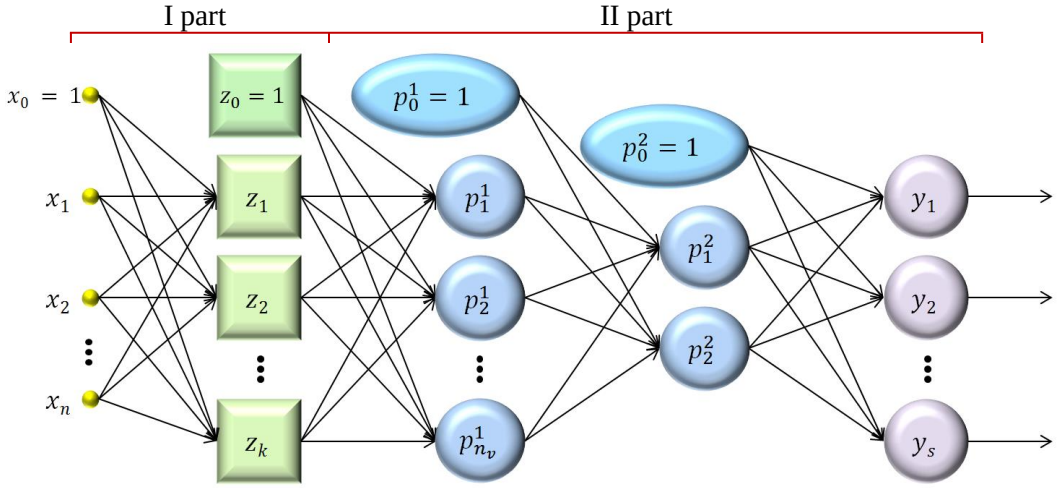


Figure 3.2 The general scheme of the network REGM

The input of the network REGM is denoted by $X = (x_1, x_2, \ldots, x_n)$. In Figure 3.2, the presented network REGM has three hidden layers of neurons. In this dissertation, the first hidden layer of neurons $Z = (z_1, z_2, \ldots, z_k)$ will be called a radial basis function layer (it is marked by green squares in Figure 3.2), while the layers of the multilayer perceptron $P^1 = (p_1^1, p_2^1, \ldots, p_{n_v}^1)$ and $P^2 = (p_1^2, p_2^2)$ will be called the first and small (or the last hidden) layer of neurons (they are marked by blue circles in Figure 3.2). The number of radial basis functions is the same as the predicted number of clusters k in multidimensional data. The number of neurons n_v in the first hidden layer $P^1 = (p_1^1, p_2^1, \ldots, p_{n_v}^1)$ can be chosen freely. There are two ($n_v = 2$) or three ($n_v = 3$) neurons in the small layer. The number of neurons depends on the space in which we wish to get the projection ($\mathbb{R}^2$ or $\mathbb{R}^3$) of multidimensional data. The output of the network REGM is denoted as $Y = (y_1, y_2, \ldots, y_s)$. The number of neurons in the output layer can be from one to k (number of the clusters). If the number of neurons in the output layer is the same as that of the clusters, then we have a certain structure similar to the "bottleneck" neural network, but the training is essentially different (see for more details in Section 3.3.). Suppose, the multidimensional data have five clusters ($k = 5$). Then there are five basis functions in the radial basis function layer, and we choose one, two, three, four or five neurons in the output layer.

We suggest to use the logical sigmoid function $f(a) = \frac{1}{1+e^{-a}}$ or linear $f(a) = a$ activation function in the hidden and output layers.

3.3. Training of the Network REGM

Before starting training the network REGM first we must set the desired network response values $T_i = (t_{i1}, t_{i2}, \ldots, t_{is})$, $i = \overline{1,m}$ (here s is the number of the neurons in the output layer) for each n-dimensional point $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1,m}$, of network input. The desired values are the centers of clusters $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $j = \overline{1,k}$, of multidimensional data. In the sequel, we present more in detail.

The available multidimensional data, expressed by the data points of the n-dimensional space $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1,m}$, are clustered into the chosen number k of clusters K_j by the k-means method. Thus, we set the centers $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, $j = \overline{1,k}$, of clusters K_j.

The cluster center μ_j bearing the input data point X_i is assigned to the desired value T_i. For this reason, there often will be identical desired values of different points of the input data. Note that $n \neq s$, therefore we propose two strategies for selecting the desired values:

1. The centers of the cluster $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, obtained by the k-means method, are projected from the space $\mathbb{R}^n$ to the low-dimensional space $\mathbb{R}^s$ by the multidimensional scaling method, $s < n$. So we obtain the projection $\mu_j^y \in \mathbb{R}^s$, $j = \overline{1,k}$, of the cluster centers $\mu_j \in \mathbb{R}^n$. The desired values are $T_i = \mu_j^y$, if $X_i \in K_j$, $i = \overline{1,m}$. Note, that the number of neurons can be from one to k (the number of clusters) in the output layer. If $s = k$ and the projection from $\mu_j \in \mathbb{R}^n$ to $\mu_j^y \in \mathbb{R}^s$, $j = \overline{1,k}$ is performed by the multidimensional scaling method, the last component μ_j^y will always be equal to 0.

2. The transformation (how to perform the transformation of multidimensional data and cluster centers is described in Section 3.3.1.) of the cluster centers $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jn})$, from the space $\mathbb{R}^n$ to space $\mathbb{R}^k$, $(k < n)$ is performed by the radial basis function. After the transformation, the cluster centers are denoted by $\mu_j^z = (\mu_{j1}^z, \mu_{j2}^z, \ldots, \mu_{jk}^z)$. If $s < k$, the transformed cluster centers μ_j^z are projected from the space R^k to the space R^s even by the multidimensional scaling method. We obtain the projections $\mu_j^y \in \mathbb{R}^s$, $j = \overline{1,k}$, of the cluster centers $\mu_j^z \in \mathbb{R}^k$. $T_i = \mu_j^y$, if $X_i \in K_j$, $i = \overline{1,m}$. Note that, if $s = k$, then $T_i = \mu_j^z$, i.e. the projection from $\mathbb{R}^k$ to $\mathbb{R}^s$ is not needed.

The training of the network REGM takes place in two stages. In Figure 3.2, each stage is marked in the scheme of the network REGM.

Stage I. The transformation of multidimensional data to a lower-dimensional space is performed by using the radial basis functions;

Stage II. The multilayer perceptron is trained by the error back propagation learning algorithm.

3.3.1. The First Stage

The transformation of multidimensional data is performed by using the exponential (3.1) or Gaussian (3.2) function in the first stage.

When transforming multidimensional data from $X_i \in \mathbb{R}^n$, $i = \overline{1,m}$ into $Z_i \in \mathbb{R}^k$, $i = \overline{1,m}$, using the exponential or Gaussian functions, it is important to choose the proper parameters of the functions: the centers μ_j and the width parameter σ. Just like most of the authors (Pierrefeu ir kt., 2006; Benoudjit ir Verleysen, 2003), in this dissertation, the centers are proposed to choose by clustering data using the k-means method. However, the results of the exponential and Gaussian functions depend not only on the proper choice of the centers μ_j but also on the width parameter σ.

The width parameter σ can be calculated according to the formula:

$$\sigma = \alpha d_{\text{avg}}, \tag{3.3}$$

where α is the constant, d_{avg} is the average distance between centers of the cluster.

The value of the constant α is determined by the scattering of objects in each cluster, i.e. the dispersion is calculated:

$$D_{K_j} = \frac{1}{km_{K_j} - 1} \sum_{X_i \in K_j} \sum_{\tilde{j}=1}^{k} \left(x_{i\tilde{j}}^{K_j} - \overline{x}_{K_j} \right)^2, \tag{3.4}$$

where K_j is the jth cluster, $j = \overline{1,k}$; k is the number of clusters; m_{K_j} is the number of objects in the cluster K_j, $\sum_{j=1}^{k} m_{K_j} = m$; $x_{i\tilde{j}}^{K_j}$ is the value of $\tilde{j}$th feature of the ith object of cluster K_j; $\overline{x}_{K_j}$ is the general average of the object feature values of cluster K_j:

$$\overline{x}_{K_j} = \frac{1}{km_{K_j}} \sum_{X_i \in K_j} \sum_{\tilde{j}=1}^{k} x_{i\tilde{j}}^{K_j}.$$

The constant α is chosen from a certain interval. This interval elapses by step 0.01 and in each iteration the value τ is calculated according to the formula:

$$\tau = \frac{1}{k} \sum_{j=1}^{k} D_{K_j}, \tag{3.5}$$

where τ is the average of dispersions (3.4).

Each obtained value τ is compared with the previously obtained value τ. As the difference between the values τ reaches the set accuracy $\epsilon = 0.0001$ ($0 < \tau^{u-1} - \tau^u \leq 0.0001$, where u is the order number of iteration), then the value of the constant α is fixed and the iterative process is stopped. The obtained value of the constant α is inserted into Formula (3.3) and the width parameter σ is calculated.

The radial basis functions become fully defined, when the proper values of the center μ_j and the width parameter σ are found.

3.3.2. The Second Stage

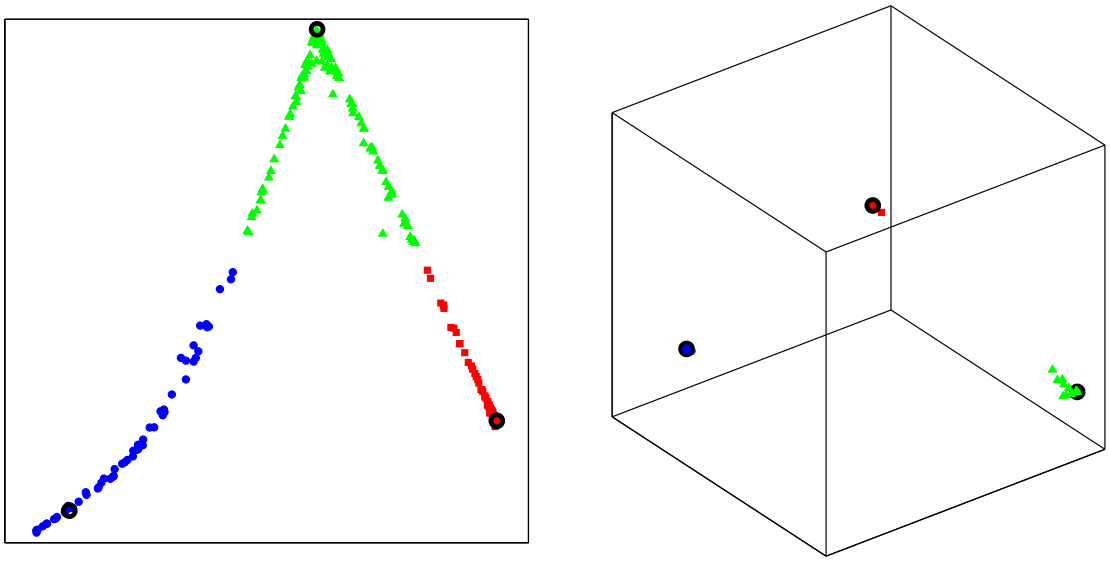
The multilayer perceptron is trained by the error back propagation learning algorithm in the second stage. The inputs of the multilayer perceptron are the new data set $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_m\} = \{z_{ij}, i = \overline{1,m}, j = \overline{1,k}\}$, obtained after the transformation of $\mathbf{X}$. The desired values T_i can be determined by one of the two ways, described in Section 3.3.

3.4. The Visualization Quality Criterion of the Results

As mentioned in Section 3.2., the projection of multidimensional data after training of the network REGM, presented in Figure 3.2, is obtained in the small layer output, when the n-dimensional analysed data set $\mathbf{X}$ is fed to the network. The projection, presented visually, should help the researcher to reveal the properties of the clusters of multidimensional data.

Note that the visual results of the output layer of the network REGM, presented in Figure 3.2, show us whether the network quality pays off. Since the desired values of the network REGM are the centers of the clusters, so in the ideal case, there should be as many different values in the output layer, as the clusters are chosen in the data. For simplicity, visualization of the results of the small and output layers, obtained after training the network REGM, in a two-dimensional or three-dimensional space will be called the obtained visualization results.

The obtained visualization results in the small layer P^2 of neurons and output layer Y, when the network REGM is trained by the Heart disease data set, are presented in Figure 3.3.



(a) Small layer **(b)** Output layer

Figure 3.3 Visualization results of the network REGM, $E(W) = 0.0006$

There is no marking of scales in figures because only intergroup positioning of points is relevant. The objects of different clusters are marked in different colours and symbols in Figure 3.3 (the first cluster – ●; the second cluster – ▲; the third cluster – ■). The cluster centers are marked by ⊙. Note that, the information about the object attribution for a specific cluster is obtained by the k-means method, which is used for setting the centers μ_j of radial basis functions. There are three rather compact accumulations of cluster points in Figure 3.3b. It illustrates a fairly good network training, because in the ideal case, there should be only three points. The Heart disease data set projection, obtained in the small layer, was compared with that obtained by the multidimensional scaling method, presented in Figure 3.4. Note that in the data projection on plane, in the case of the multidimensional scaling method, it is tried to sustain the distances between the points before and after the projection. Meanwhile, the distances between the points in the network REGM are not attempted to sustain, but only to stress the intergroup similarities/differences.

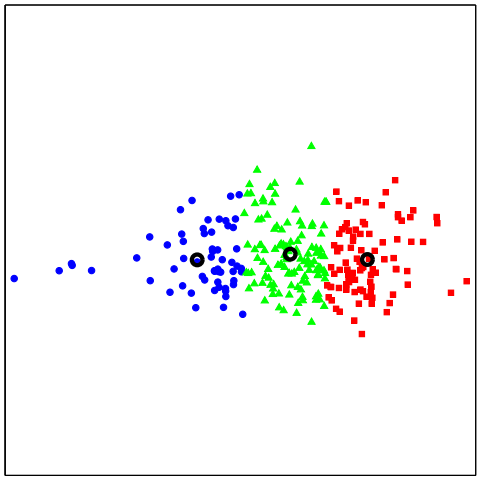


Figure 3.4 Projection obtained by the multidimensional scaling method

There are a lot of points that make up the cluster in both Figures 3.3a and 3.4, i.e. the points in the cluster are not swarmed around their cluster center as we see in the visualization results of the output layer in Figure 3.3b. However, the distribution of the points in the clusters is different. In the projection, obtained by the multidimensional scaling method, the points of each cluster are scattered in all directions from the center of the cluster (Figure 3.4). In the projection, obtained after the network REGM training, the points in the cluster are located around the several straight lines or curves. The advantage of such a location around the several straight lines or curves is that the points which have some similarities with that of the neighbouring cluster are exposed or the points which are specific to a particular cluster are distinguished. The points which have the similarities with that of the neighbour cluster sometimes are even separated from the rest of the cluster points. For example, two separate groups of points stand out in the cluster marked ■ in Figure 3.3a. The group nearest to the cluster marked ▲ has more similarities with these cluster points than the further group. In Figure 3.3a, the boundaries between the clusters are obvious, while no clear boundary (gap) is seen between the clusters in Figure 3.4.

Not all data projections presented visually are informative and answer the set purpose of the dissertation. Therefore, the visualization quality criteria for the projection of the data in the small layer where set that evaluate the obtained visualization result:

1. distribution of points around the straight lines or curves;
2. scattering of points in the cluster;
3. the boundary between clusters.

The first visualization quality criterion is qualitative and the other two are quantitative.

The first visualization quality criterion refers to point distribution in the projection. The cluster of points in the visualized data should make up straight lines or curves. Such a distribution reveals their intergroup similarities and differences. The clusters of points that have similarities with only one neighbouring cluster are distributed around the one straight line or a curve. The clusters of points in the

best case having similarities with the some of neighbouring clusters are distributed around the some straight lines or curves.

The second visualization quality criterion is closely connected with the first visualization quality criterion. This criterion shows that, in each cluster, as many points, forming a cluster, as possible should be seen. Since the points in the projection are distributed around the straight lines or curves, the point scattering on the straight lines or curves can be calculated according to the maximum distance between the points of the cluster K_j, $j = \overline{1,k}$. The distance is denoted by $\bar{a}_{K_j}$. Only the projections, whose maximum distances $\bar{a}_{K_j}$ between clusters are larger than 0.1 ($\bar{a}_{K_j} > 0.1$), meets for the second visualization quality criterion.

The first two visualization quality criteria are most important. The third visualization quality criterion is desirable, but not obligatory. This criterion states that there must be a boundary between clusters, i.e. a certain gap between different clusters. The third visualization criterion is the minimum distance between the points of adjacent clusters and it is denoted by $\hat{a}$. Only the projections, in which the minimal distance $\hat{a}$ between the points of adjacent clusters is equal to or lager than 0.05 ($\hat{a} \geq 0.05$), meets for the third visualization quality criterion.

Two selection criteria are proposed in this dissertation:

1. Saving of clusters in the data after network training.
2. Scattering of the points obtained in the output layer.

The value of the criterion of saving of clusters in the data should reflect how much s-dimensional points have changed dependencies to the clusters as compared to the n-dimensional case.

One of the cluster characteristic is its center. The cluster weight centers μ_j, $j = \overline{1,k}$, of data set $\mathbf{X}$ are used in radial basis functions. The network is trained by data set $\mathbf{X}$. The cluster weight centers μ_j, $j = \overline{1,k}$, are presented to the network and the s-dimensional projections μ_j^y, $j = \overline{1,k}$, of the n-dimensional centers are obtained in the output. Clustering of the s-dimensional points Y_i, $j = \overline{1,m}$, is performed as follows:

1. Y_i, $i = \overline{1,m}$, are calculated by giving X_i, $i = \overline{1,m}$ into the network.
2. The distances between s-dimensional Y_i and s-dimensional centers μ_j^y, obtained at the output, are calculated.
3. The point Y_i is assigned to the cluster K_j^y, the distance to the center μ_j^y of which is minimal.

In the ideal case, the clusters K_j^y and K_j should consist of the points marked by the same numbers, i.e if $X_i \in K_j$, then $Y_i \in K_j^y$ as well. However, in the general case, it may occur that $X_i \in K_j$, and $Y_i \notin K_j^y$.

The value of the criterion χ of saving of clusters in the data is the total number of the s-dimensional points Y_i in all the clusters K_j^y, $j = \overline{1,k}$, where the following condition is valid: $Y_i \notin K_j^y$, as $X_i \in K_j$.

After evaluating the cases by the first selection criterion, the points that do not satisfy the condition $Y_i \in K_j^y$, as $X_i \in K_j$ are rejected, and we go on to calculation of the second selection criterion, using the rest set of points $\{\tilde{Y}_{\hat{i}}, \hat{i} = \overline{1, m-\chi}\}$. Note that if $\chi = 0$, then $\{\tilde{Y}_{\hat{i}}, \hat{i} = \overline{1, m-\chi}\} = \{Y_i, i = \overline{1,m}\}$.

After rejecting the network training results by the first selection criterion, there remain $\widetilde{c}$ the results of the network training, where the value χ is the least one. If $\widetilde{c} > 1$, then we have to use the second selection criterion.

The intergroup location of points of different clusters $\tilde{Y}_i \in K_j^y$, as $X_i \in K_j$, can be evaluated by calculating the distance between points of different clusters. Denote the maximum distance between points of different clusters in all $\tilde{c}$ trains of the network by κ:

$$\kappa = \max_{q=\overline{1,\widetilde{c}}} \kappa_q, \tag{3.6}$$

where κ_q is the minimum distance between the points of different clusters in the qth network training, calculated according to the formula:

$$\kappa_q = \min_{1 \le j_1 < j_2 \le k} \min_{\substack{\tilde{Y}_{i_1} \in K_{j_1}^y \\ \tilde{Y}_{i_2} \in K_{j_2}^y}} \|\tilde{Y}_{i_1} - \tilde{Y}_{i_2}\|. \tag{3.7}$$

Here $\tilde{Y}_i$, $i = \overline{1, m-\chi}$, are the values obtained after the qth network training. For simplicity, the index q on the right side of Formula (3.7) is not introduced.

If the value κ_q is large, it means that the points $\tilde{Y}_i \in K_j^y$ are close to their cluster centers and in the visual results, accumulation of the points will be clearly visible. So, if we have $\widetilde{c}$ different κ_q values, $q = \overline{1,\widetilde{c}}$, then the best result will be in that network training q, where κ_q is maximal.

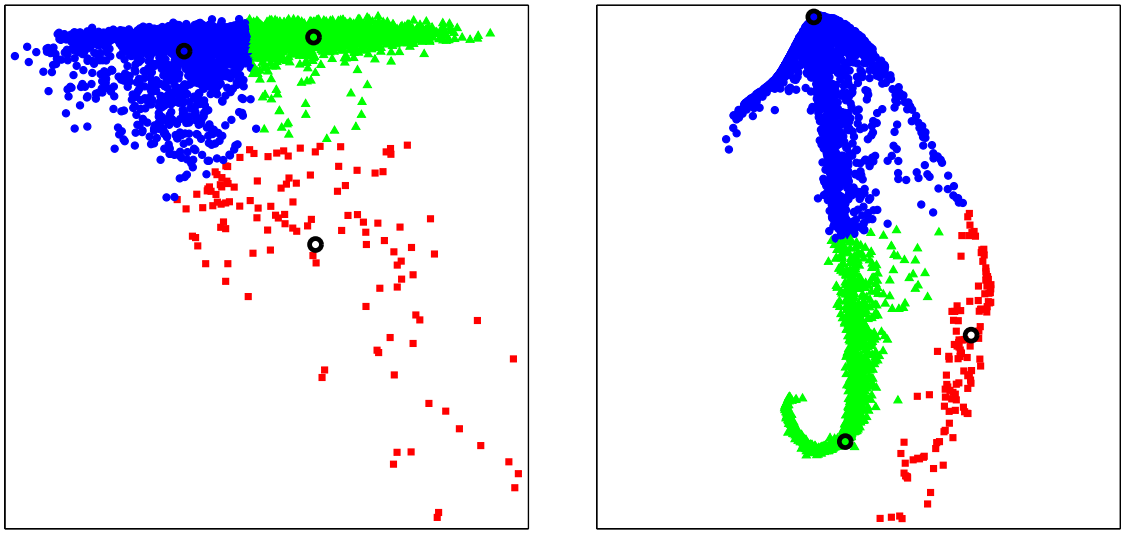
The values of the weights of the network are fixed, if the trained network meet for both selection criteria (i.e. clusters are saved in the data and the value of κ_q is maximal in $\widetilde{c}$ network training). After adding the new objects to the data set and giving them to the network REGM with the fixed values of weights, their place on the projection is displayed without complicated calculations (network retraining).

3.5. The Practical Use of the Network REGM

The network REGM is used for visual analysis of multidimensional data. The values, displayed in a two-dimensional or three-dimensional space, are obtained in the last hidden layer of neurons, when the n-dimensional data set $\mathbf{X}$ is filed to the network. The number of groups of the points that make up the data set $\mathbf{X}$ is not known. The classical clustering methods classify data into groups but they do not reveal intergroup similarities/differences of the objects. Meanwhile, the data projection on the plane, obtained by the network REGM, reflects the general structure of the data (clusters, proximity between clusters, intergroup similarities of points) rather than interlocation of multidimensional points.

Let us consider an example of the Wilt data set. The problem is to identify and isolate the withered and starting to wilt trees. The data, clustered by the k-means method and visualized by the multidimensional scaling method are presented in Figure 3.5a. The objects of different clusters are marked in different colours: the healthy trees are marked ●, wilt trees are marked ■, the remaining land surface is marked ▲. From Figure 3.5a we can say which object belongs to which cluster, but it is very difficult to assess intergroup similarities/differences of objects. It is of great importance, because if we can find the trees which have similarities with the wilt trees, we would be able to establish the reason of tree wilting (deficiency of moisture

or disease). The projection, obtained after the training network REGM presented in Figure 3.5b, allow us to assess intergroup similarities/differences of objects.



(a) MDS **(b)** REGM

Figure 3.5 Projections of the Wilt data set

The network REGM, presented in Figure 3.2, was trained 60 times. The best projection of the network was found according to the selection criteria (Figure 3.5b). This projection was evaluated by the set visualization quality criteria. The first visualization quality criterion shows us that the points must be distributed around the straight lines or curves. We see in Figure 3.5b that the points are scattered around the three curves. So, the first visualization quality criteria is satisfied. The second visualization quality criterion shows us scattering of the points in the cluster: the maximum distance between points in the cluster denoted by ●, is $\bar{a}_{K_1} = 0.48$, in the cluster denoted by ▲, $\bar{a}_{K_2} = 0.51$, and in the cluster denoted by ■, $\bar{a}_{K_3} = 0.62$. We see that all the values $\bar{a}_{K_j}$ are larger than 0.1, so the second visualization quality criterion is also satisfied. The third visualization quality criterion (the boundary between clusters) is desirable, but not obligatory. In this case, the distance $\hat{a}$ between the clusters marked ● and ▲, is 0.006, and between the clusters marked ● and ■, is $\hat{a} = 0.02$. The distance $\hat{a}$ between the clusters should be larger or equal to 0.05. The projection presented in Figure 3.5b does not satisfy the third visualization quality criterion.

We can see in Figure 3.5b that the objects of the cluster, marked ● are as if divided into three groups. The objects of each group are scattered around the separate straight lines or curves. The objects of healthy trees (the cluster is marked ●), which have similarities with the objects of wilt trees (the cluster is marked ■), are scattered on the curve nearest to the wild trees cluster. Namely to these objects we should pay attention, because they can help the researcher to determine the cause of tree wilting. From the projection, presented in Figure 3.5b, we can state that the objects of the wild trees cluster have no similarities with the objects of the remaining land surface cluster, because there is no connecting straight line or curve.

4. Experimental Research

4.1. Network REGM Used in the Experiments

The experiments were done with the hybrid neural network REGM presented in Figure 3.2. Three clusters were selected in the data, therefore in the radial basis function layer Z, there are three radial basis functions. An exponential radial basis function was used in this layer. The Gaussian radial basis function was only used in the experiment "Selection of the desired values". The number of neurons in the first hidden layer P^1 was chosen equal to five. The number of neurons in the small layer P^2 was chosen equal to two, because after network training it is desirable to visualize data on the plane. The number of neurons in the output layer Y was chosen equal to the number of the selected clusters in multidimensional data. The number of neurons was chosen from one to k in the experiment "The number of neurons in the output layer". The linear activation function was used in the small layer. The logical sigmoid activation function was used in the first hidden and output layers. Different activation functions were used only in the experiment "Activation functions of the second part of the network REGM".

4.2. Selection of the Desired Values

The desired values are very important in the experiments with the network REGM, because the efficiency of the trained network REGM performance in data visualization depends on them. Two strategies are offered to choose the desired values $T_i = (t_{i1}, t_{i2}, \ldots, t_{is})$, $i = \overline{1, m}$ as mentioned in Section 3.3.

For simplicity, let us call the first strategy of selecting desired values by *untransformed centers*, and the second strategy by *transformed centers*.

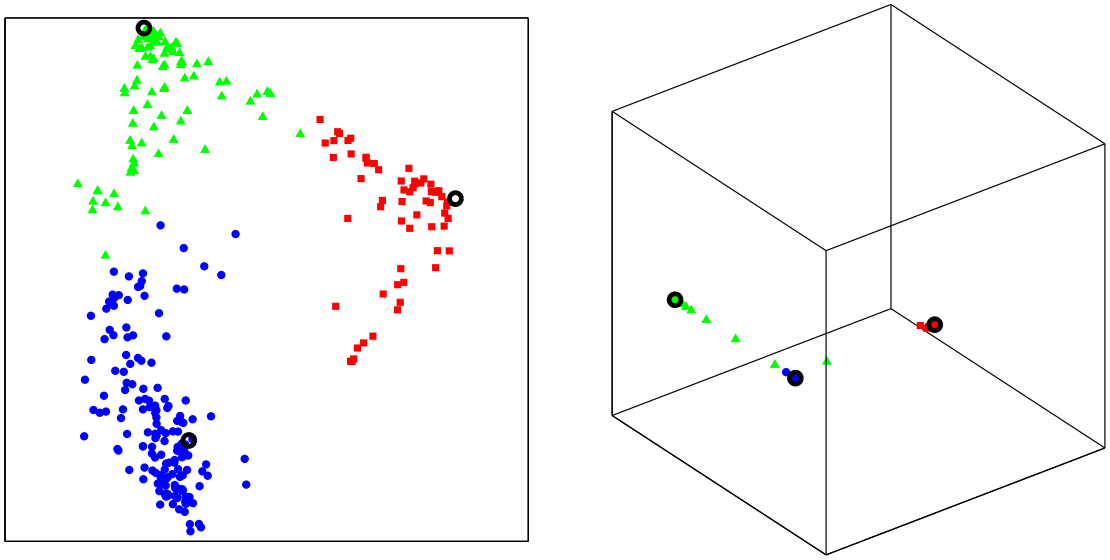
The aim of the experiment is to define the appropriate strategy for desired values by which the network REGM is trained with a higher quality and the visualization results in the small layer meet for the quality criteria set for visualization (see Section 3.4).

Both in untransformed and transformed cases, the network has been trained for 20 times in the experiment. The same 20 initial sets of weights were used in both cases. That enables us to compare both strategies without using a large amount of long calculations. The results of the experiment illustrated Vertebral column data set.

The visualization results of small and output layers are presented in Figure 4.1. The second and the third visualization quality criteria values of the small layer are presented in Tables 4.1 and 4.2.

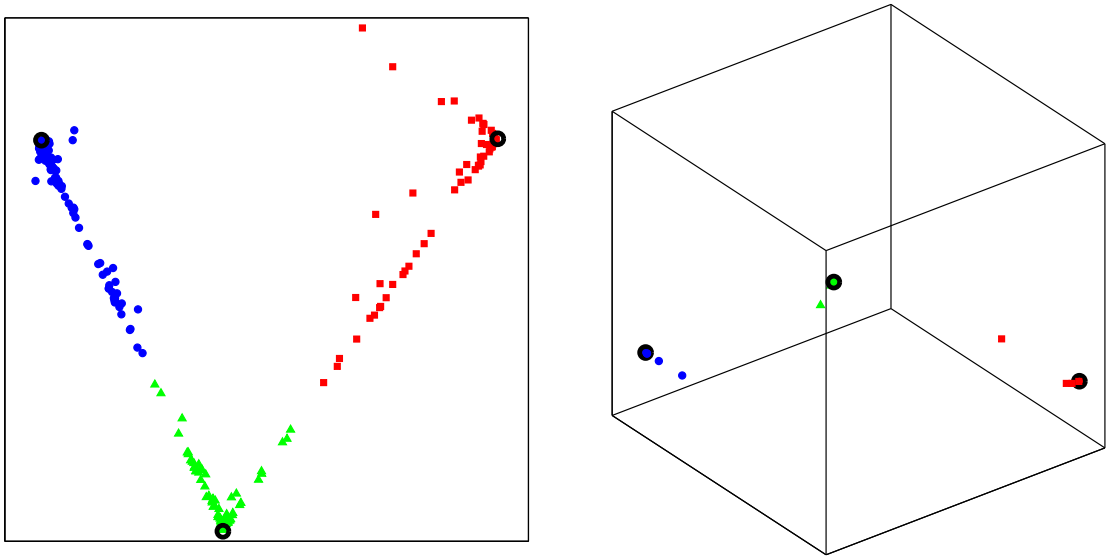
Table 4.1: Values of the second visualization quality criterion of Figure 4.1

Figure	● cluster	▲ cluster	■ cluster
4.1a	0.61	0.46	0.48
4.2a	0.45	0.31	0.68



(a) Small layer **(b)** Output layer

Figure 4.1 Desired values in the case of untransformed centers, $E(W) = 0.0484$



(a) Small layer **(b)** Output layer

Figure 4.2 Desired values in the case of transformed centers, $E(W) = 0.0009$

Table 4.2: Values of the third visualization quality criterion of Figure 4.1

Figure	● and ▲ clusters	▲ and ■ clusters
4.1a	0.03	0.04
4.2a	0.07	0.11

The network makes less errors after training when the desired network respond values are transformed centers. We see the cluster points scattering in the visualization results of the output layer. In the small layer, the cluster points are situated "in clouds", but not around the straight lines or curves. The estimates, presented in Tables 4.1 and 4.2, show that the visualization results meets for only to the second visualization quality criterion.

We see only three points in the output layer, because there are only three clusters in the data set. The visualization results conform to all the three visualization quality criteria in the small layer.

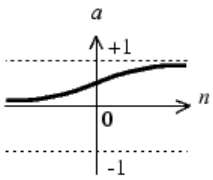
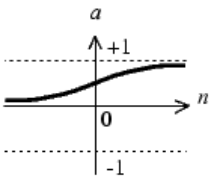
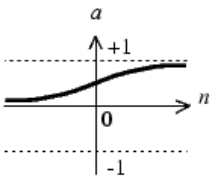
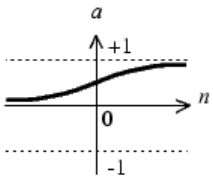
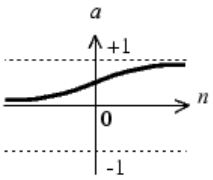
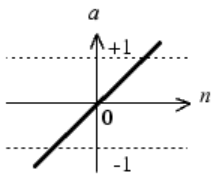
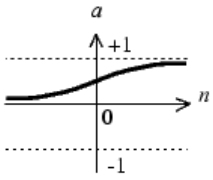
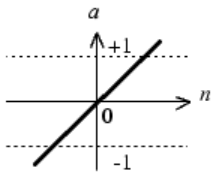
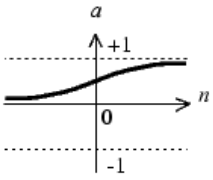
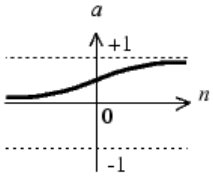
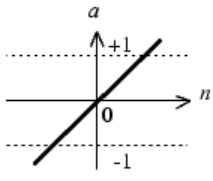
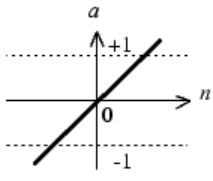
Summarizing the results of the experiment we can conclude that the quality of the training of hybrid neural network REGM is higher and the visualization results meet for the set visualization quality criteria if the transformed centers are taken as the desired values and the exponential function is used in the radial basis function layer.

4.3. Activation Functions of the Second Part of the Network REGM

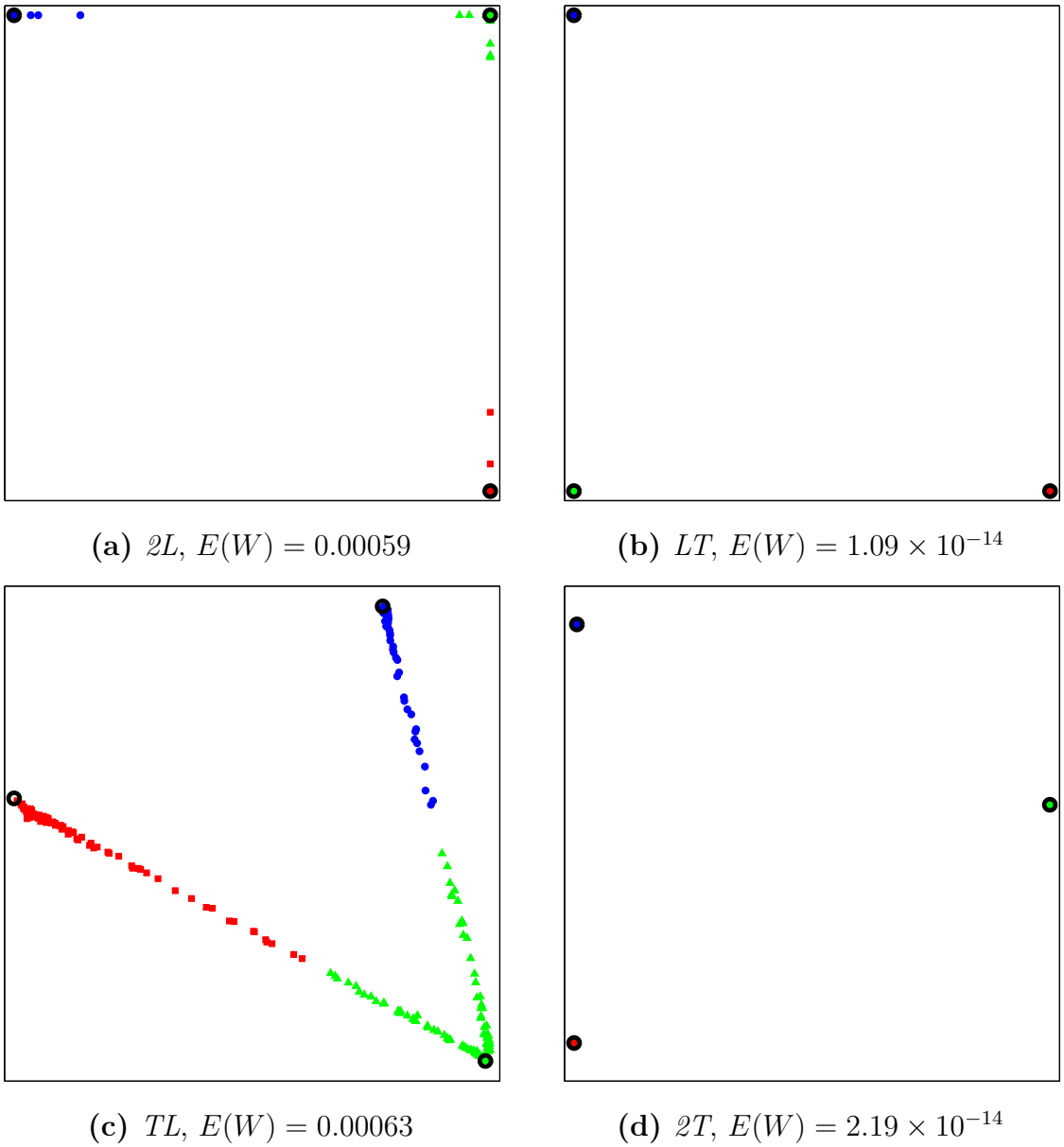
Another very important factor of the network REGM training is activation functions, present in the hidden and output layers in the second part of the network REGM. As mentioned in Section 3.2. the logical sigmoid or linear activation functions are used in these layers. There arises only one question, in which layer and which activation function it is best to use, so that the visualization result after the network REGM training would satisfy the set quality criteria of visualization.

Four experiments were carried out. Different activation functions were used in small and output layers during the experiments. The use of activation functions in the experiments is presented in Table 4.3. For simplicity, let us denote the first experiment by *2L*, the second by *LT*, the third by *TL* and the fourth one by *2T*. The logical sigmoid activation function was used in the first hidden layer. The network was trained 30 times in each experiment.

Table 4.3: Selection of the activation function in the hidden and output layers

	P^1	P^2	Y
2L			
LT			
TL			
2T			

The results of the experiment are illustrated by the Heart disease data set and are presented in Figure 4.3.



(a) $2L$, $E(W) = 0.00059$ **(b)** LT, $E(W) = 1.09 \times 10^{-14}$

(c) TL, $E(W) = 0.00063$ **(d)** $2T$, $E(W) = 2.19 \times 10^{-14}$

Figure 4.3 Visualization results in the small layer after training

Values of the second and third visualization quality criteria are presented in Tables 4.4 and 4.5.

Table 4.4: Values of the second visualization quality criterion

Figure	● cluster	▲ cluster	■ cluster
4.3a	0.09	0.08	0.12
4.3b	0.000009	0.000064	0.000003
4.3c	0.38	0.39	0.61
4.3d	0.00003	0.00033	0.00041

Table 4.5: Values of the third visualization quality criterion

Figure	● and ▲ clusters	▲ and ■ clusters
4.3a	0.56	0.53
4.3b	0.70	0.70
4.3c	0.09	0.06
4.3d	0.95	0.99

It is very important that the small layer visualization result obtained in the experiment *2L* does not meet for the second visualization quality criterion. There are only as many points as there are clusters in the data in the small layer visualization results, obtained in the experiments *LT* and *2T*. These images provide no more knowledge than the image in the output layer. The visualization results are informative and meet for all the three visualization quality criteria in the small layer in the experiment *TL*:

1. The points are situated around the the straight lines or curves;
2. The points are scattered in clusters (the largest distances between cluster points are larger than 0.1);
3. The bounds between clusters (the smallest distance between the points of different clusters is larger than 0.05).

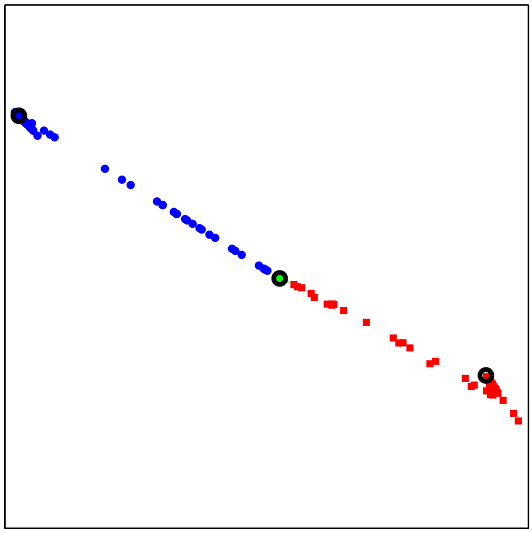
Summarizing the results of four experiments we can conclude that the quality of the training of hybrid neural network REGM is higher and the visualization results meet for the set visualization quality criteria if the linear activation function is used in the small layer and the logical sigmoid activation function is used in the first hidden and output layers.

4.4. The Number of Neurons in the Output Layer

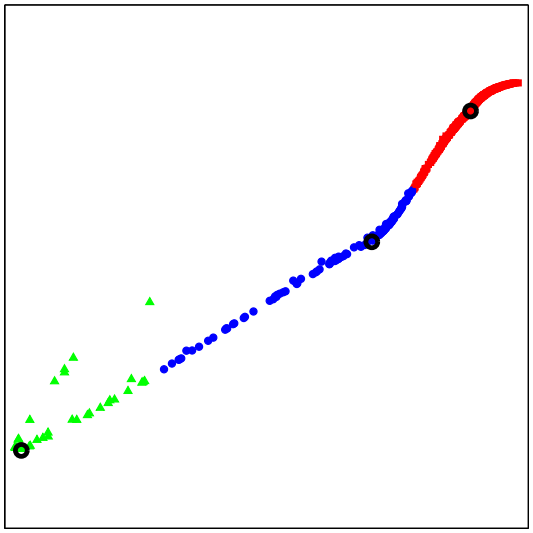
In this experiment with the hybrid neural network REGM, we observed the evolution of visualization results obtained in the small layer, when different number of neurons in the output layer is chosen. The number of neurons can be from one to k in the output layer. In the first experiment, it was $s = 1$; in the second experiment $s = 2$; in the third experiment $s = 3$. The desired network response values are transformed cluster centers. Let us note that the transformed cluster centers $\mu_j^z \in \mathbb{R}^k$ were projected to the space $\mathbb{R}^s$ by the multidimensional scaling method in the first and second experiments, because $s < k$, while the projection of transformed data is unnecessary in the third experiment because $s = k$. The network was trained 30 times in one experiment.

The visualization results of the experiment were illustrated by data sets of Vertebral Column, Breast Cancer, and Heart diseases.

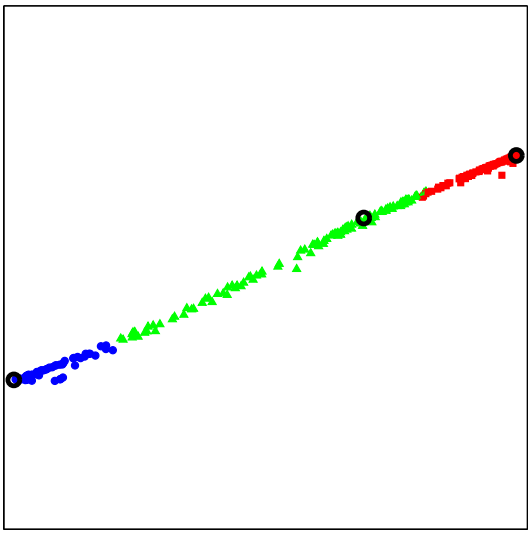
The values of the second and third visualization quality criteria are presented in Tables 4.6, 4.7, and 4.8 of Figures 4.4, 4.5, and 4.6.



(a) Vertebral Column data set



(b) Breast Cancer data set



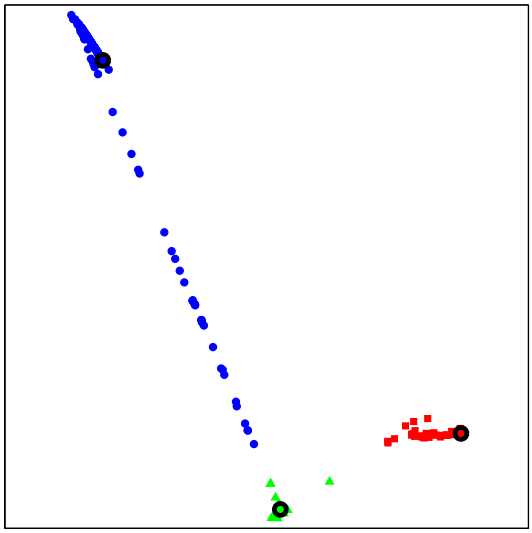
(c) Heart diseases data set

Figure 4.4 REGM training results when one neuron is chosen in the output layer

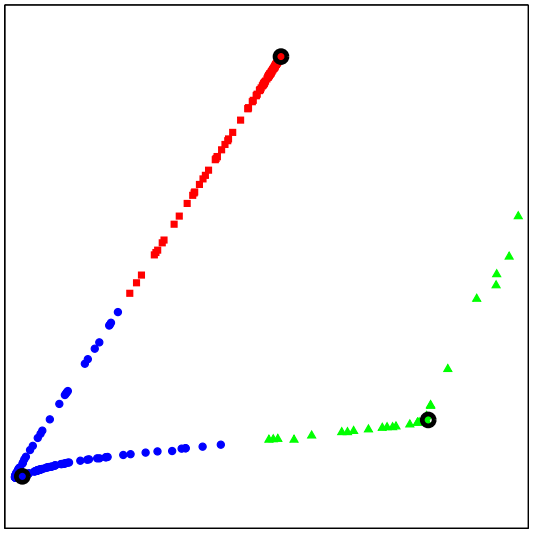
Table 4.6: Values of the second and third visualization quality criteria for Vertebral Column data set

Output layer	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● and ▲	▲ and ■
$s = 1$	0.50	0.01	0.44	0.02	0.02
$s = 2$	0.82	0.12	0.13	0.08	0.12
$s = 3$	0.52	0.35	0.37	0.09	0.08

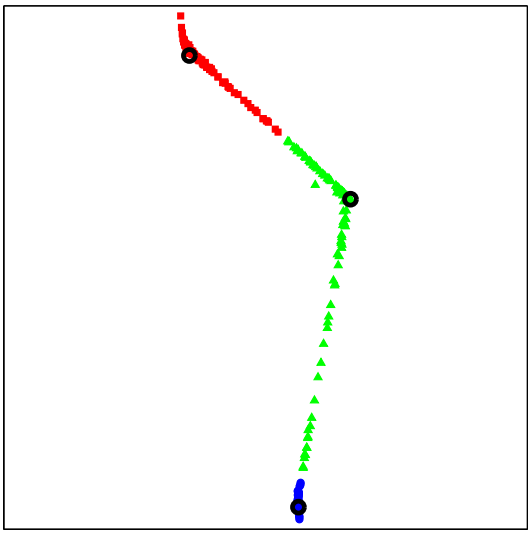
The first visualization quality criterion does not meet for visualization results in Figure 4.4, because the located points should be around the several straight lines or curves. If the quality of the first visualization criterion is not satisfied, then the other visualization quality criteria can be disregarded.



(a) Vertebral Column data set



(b) Breast Cancer data set



(c) Heart diseases data set

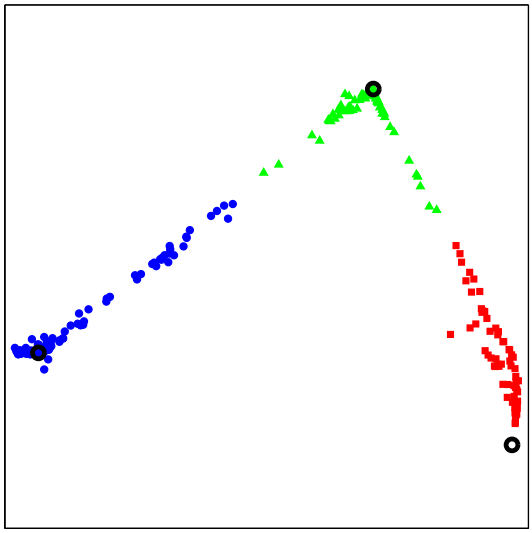
Figure 4.5 REGM training results when two neurons are chosen in the output layer

Table 4.7: Values of the second and third visualization quality criteria for Breast Cancer data set

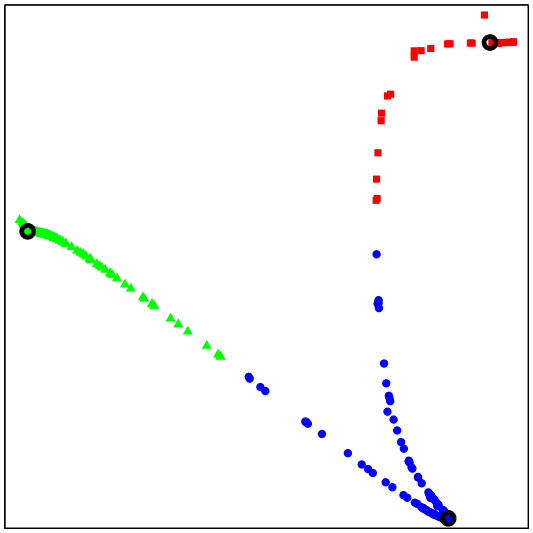
Output layer	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● and ▲	● and ■
$s = 1$	0.49	0.32	0.24	0.04	0.003
$s = 2$	0.37	0.59	0.49	0.09	0.04
$s = 3$	0.52	0.46	0.41	0.07	0.10

The first visualization quality criterion meets for visualization results in Figure 4.5. The second visualization quality criterion does not meet only for the Heart disease data set, as we see in Tables 4.6, 4.7, and 4.8. The third visualization quality criterion meets for only Vertebral Column data set.

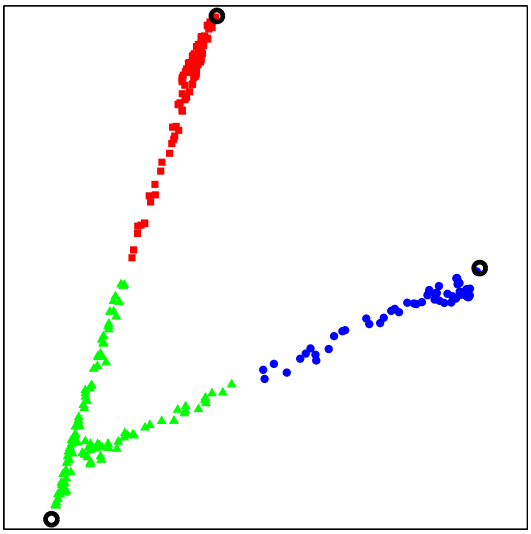
All the three visualization quality criteria meet for the visualization results of all the data sets (see Figure 4.6 and Tables 4.6, 4.7, 4.8).



(a) Vertebral Column data set



(b) Breast Cancer data set



(c) Heart diseases data set

Figure 4.6 REGM training results when three neurons are chosen in the output layer

Table 4.8: Values of the second and third visualization quality criteria for Heart data set

Output layer	$\bar{a}_{K_j}$			$\hat{a}$	
	●	▲	■	● and ▲	▲ and ■
$s = 1$	0.18	0.62	0.19	0.03	0.003
$s = 2$	0.07	0.63	0.29	0.03	0.02
$s = 3$	0.46	0.45	0.49	0.06	0.05

Summarizing the results of experiments we can conclude that the quality of the training of hybrid neural network REGM is higher and the visualization results meet for the set visualization quality criteria if the number of neurons in the output layer is equal to the number of clusters selected in the data set, i.e. $s = k$.

5. Summary and General Conclusions

The analytical survey of hybrid neural networks (various radial basis functions and multilayer perceptron compounds) has showed that networks of this type are constructed in many different fields and for solving specific tasks. The results of hybrid networks are more accurate in comparison with the results of radial basis function neural networks or multilayer perceptrons. The hybrid neural network architecture for solving a specific task is chosen according to the individual characteristics of separate networks.

A new hybrid neural network REGM is proposed and investigated. This neural network integrates the ideas both of the radial basis function neural network and that of a multilayer perceptron, which has the properties of a "bottleneck" neural network. The network REGM consists of two parts: the radial basis function layer and the multilayer perceptron of the special structure. The first part is a kind of transformation of multidimensional points into the space of desired lower dimensionality. The second part is a multilayer perceptron with the last hidden layer (so-called small layer) composed of a small number of neurons (2 or 3). The purpose of the network REGM is to help in discovering visually the properties of clusters in the multidimensional data set.

The network REGM is used for the visual analysis of multidimensional data in such a way that the output values of the neurons of the last hidden layer are the two-dimensional or three-dimensional projections of the multidimensional data, when the analyzed data set is given to the network. A peculiarity of the network is that the visualization results on the plane reflect the general structure of the data (clusters, proximity between clusters, intergroup similarities of points) rather than the location of multidimensional points.

The research completed in this dissertation has led to the following conclusions:

1. The network REGM is a powerful new tool for visual exploration of multidimensional data because there arises an opportunity to know the overall structure of the data better. Multidimensional data clustering results can be used not only in the calculation of radial basis function parameters, but also for the visual presentation of the results on the plane.

2. If the radial basis function (RBF) width parameter is calculated according to the scattering of objects in the clusters and the average distance between their centers, then the obtained values of RBF distribute in the interval $[0, 1]$ (i.e they do not focus on the borders of this interval).

3. After training the network REGM for several times, the proposed two criteria facilitate selection of the best projection of the data set. Using them the selection can be automated:
   - criterion of saving of clusters in the data after network training is an integer and in the ideal case, is equal to 0;
   - criterion of scattering of the points obtained in the output layer is the maximum distance between the points belonging to different clusters.

4. The multidimensional data projection obtained on the small layer is evaluated by three visualization quality criteria:
   a) The points are situated around the straight lines or curves;

b) The points are scattered in the clusters (the largest distances between the cluster points are larger than 0.1);

c) The bounds between clusters (the smallest distance between the points of different clusters is larger than 0.05).

5. All the experiments with the network REGM in the dissertation have been carried out, using real data sets of practical importance. The volume of these data sets reaches 4500 objects. In the obtained projections, not only the clusters of data are visible, but also intergroup object similarities/differences may be evaluated visually. The objects from different clusters but having similarities help a researcher to pay attention to the possible significant changes of properties of the objects (for example, an early stage of disease, similarities of types of wheat, etc.) or to look for the reasons which lead to changes.

6. The quality of the training of hybrid neural network REGM is higher and the visualization results meet for the set visualization quality criteria if:
   - the transformed centers are taken as the desired values and the exponential function is used in the radial basis function layer;
   - the linear activation function is used in the small layer and the logical sigmoid activation function is used in the first hidden and output layers;
   - the number of neurons in the output layer is equal to the number of clusters selected in the data set.

List of Literature Referenced in this Summary

Benoudjit, N. and M. Verleysen (2003). On the kernel widths in radial-basis function networks. *Neural Processing Letters 18*(2), 139–154.

Borg, I. and P. Groenen (2005). *Modern Multidimensional Scaling: Theory and Applications, 2nd edn.* Springer, New York.

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

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter 11*(1), 10–18.

Karbauskaitė, R. and G. Dzemyda (2006). Multidimensional data projection algorithms saving calculations of distances. *Information Technology and Control 35*(1), 57–64.

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

Podpečan, V., M. Zemenova, and N. Lavrač (2012). Orange4WS environment for service-oriented data mining. *The Computer Journal 55*(1), 82–98.

List of Publications on Topic of Dissertation

1. **Ringienė, L.**, Dzemyda, G. Daugiamačių duomenų požymių mažinimas naudojantis eksponentine koreliacine funkcija. *Jaunųjų mokslininkų darbai.* Vilnius: Vilniaus universitetas. ISSN 2029-9958. 2013, Nr. 1, p. 152–158.

2. **Ringienė, L.**, Dzemyda, G. Multidimensional data visualization based on the exponential correlation function. *Baltic Journal of Modern Computing.* Riga: University of Latvia. ISSN 2255-8942. 2013, Vol. 1, No. 1, p. 9–28.

3. **Ringienė, L.**, Dzemyda, G. Specialios struktūros daugiasluoksnis perceptronas daugiamačiams duomenims vizualizuoti. *Informacijos mokslai.* ISSN 1392-0561. 2009, T. 50, p. 358–364.

About the Author

Laura Ringienė was born on the 31$^{\text{th}}$ of August, 1982 in Vilnius, Lithuania.

In 2001, she graduated from Trakai Vytautas Magnus Gymnasium. She received a Bachelor's degree in Education and Teacher training from Vilnius Pedagogical University in 2005. She received a Bachelor's degree in Informatics and Teacher training from Vilnius Pedagogical University in 2006, and Master's degree in Informatics from Vilnius Pedagogical University in 2008. From 2008 till 2013 she was a PhD student of Vilnius University, Institute of Mathematics and Informatics.

Since 2005 Laura Ringienė has been working as an information technology teacher at the Vilnius Vytautas Magnus Gymnasium. Since 2008 Laura Ringienė has been working as an engineer at the Vilnius University Institute of Mathematics and Informatics.

HIBRIDINIS NEURONINIS TINKLAS DAUGIAMAČIAMS DUOMENIMS VIZUALIZUOTI

Tyrimų sritis

Sparčiai vystantis šiuolaikinėms technologijoms labai didėja kaupiamų duomenų apimtys įvairiose srityse: technikoje, ekonomikoje, medicinoje, ekologijoje ir daugelyje kitų. Duomenys kaupiami tam, kad vėliau iš jų būtų galima gauti naujų žinių, pavyzdžiui, prognozuoti būsimą veiklą, identifikuoti kritinius atvejus, apibendrinti. Tačiau turimus labai didelės apimties duomenis (dažniausiai vadinamus daugiamačiais duomenimis) žmogui savarankiškai suvokti ir interpretuoti labai sudėtinga. Tam tikslui yra kuriami įvairūs duomenų tyrybos metodai, kurie sprendžia įvairius uždavinius: suskirsto duomenis į grupes, nustato duomenų struktūrą, randa tarpusavio ryšius ar net išskirtinumus, ir pan. Čia paminėtų uždavinių sprendimą padeda (palengvina) surasti daugiamačių duomenų vizualizavimas dvimatėje arba trimatėje erdvėje. Šio darbo tyrimų sritis yra duomenų tyryba remiantis daugiamačių duomenų vizualia analize.

Darbo aktualumas

Šioje disertacijoje tiriami tokie daugiamačiai duomenys, kurie aprašo objektų (žmonių, įrenginių, augalų, gamtos reiškinių ir kt.) rinkinius, kuriuos charakterizuoja tam tikri skaitiniai požymiai (parametrai, savybės). Objektų, sudarančių konkretų analizuojamą duomenų rinkinį, skaičius m yra baigtinis. Tam tikras požymių reikšmių rinkinys nusako vieną konkretų analizuojamo duomenų rinkinio objektą $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = \overline{1,m}$, čia n yra požymių skaičius, i yra objekto numeris. Objektai X_i dar gali būti interpretuojami kaip n-mačiai taškai, o požymiai $x_1, x_2, \ldots, x_n$ – taškų koordinatėmis. Analizuojamų duomenų rinkinį galima aprašyti kaip matricą $\mathbf{X} = \{X_1, X_2, \ldots, X_m\} = \{x_{ij}, i = \overline{1,m}, j = \overline{1,n}\}$, kurios i-oji eilutė yra n-matės euklidinės erdvės taškas $X_i \in \mathbb{R}^n$ (Dzemyda ir kt., 2013).

Daugiamačių duomenų vizualizavimui jau yra sukurta nemažai metodų, bet jie ir toliau sparčiai vystomi siekiant lengvinti duomenų interpretavimą ir suvokimą (Dzemyda ir kt., 2013). Taip pat šie metodai yra realizuoti daugelyje programų sistemų: Matlab (R2009b, The MathWorks, http://www.mathworks.se/), Orange (Podpečan ir kt., 2012), Weka (Hall ir kt., 2009), ir kt. Vizualizavimo metodai turimus daugiamačius duomenis pateikia žmogui suvokiamoje erdvėje (dvimatėje arba trimatėje) perteikiant taškų išsidėstymą, t. y. išlaikant jų panašumus ir skirtingumus. Tačiau atsiranda poreikis vizualiai įvertinti duomenų rinkinio struktūrą ir savybes: susidariusias grupes, žymiai išsiskiriančius objektus, objektų panašumus/skirtingumus, ir pan. Retame duomenų rinkinyje aiškiai atsiskiria objektų grupės, t. y. matoma riba tarp objektų grupių, kaip pateikta 1.1a paveiksle, kuriame į plokštumą vizualizuotas E.coli bakterijų duomenų rinkinys (angl. *ecoli data set*). Duomenų rinkinį sudaro 336 bakterijos, kurios apibūdintos 7 požymiais. Matome tris objektų grupes, nors praktiškai jų yra daugiau. Dažniausiai skirtingų objektų grupės yra susiglaudusios arba net vienos grupės objektai pakliūna tarp kitos grupės objektų. Kaip pavyzdys pateikiamas Kviečių grūdų duomenų rinkinys (angl. *wheat seeds data set*) vizualizuotas į plokštumą 1.1b paveiksle. Duomenų rinkinį sudaro 210 kviečių grūdų, kurie apibūdinti 7 požymiais. Vaizdumo dėlei skirtingų grupių objektai pavaiz-

duoti skirtingomis spalvomis. Atsiranda poreikis atskirti vieną grupę nuo kitos arba išskirti objektų grupeles, kurios reikalauja nuodugnesnio tyrimo. Pavyzdžiui, gali kilti poreikis kiekvienoje Kviečių grūdų grupėje išskirti grūdus, kurie turi daugiausia panašumo su kitos grupės grūdais arba atvirkščiai – išgryninti konkrečios grupės grūdus.

Duomenų rinkinį papildžius nauju objektu ir norint jį pridėti turimame paveiksle, tenka arba iš naujo rasti visų duomenų projekcijas plokštumoje, jei duomenų vizualizavimas buvo atliktas klasikiniais vizualizavimo metodais, arba naudoti tam (naujų taškų atidėjimui) skirtus metodus, kurie yra netikslūs (pavyzdžiui, triangulіacijos metodas (Karbauskaitė ir Dzemyda, 2006)). Naujus objektus atitinkančių taškų atidėjimui plokštumoje sėkmingai taikomi ir dirbtiniai neuroniniai tinklai.

Darbo tikslas ir uždaviniai

Disertacijos tikslas yra sukurti metodą tokios duomenų projekcijos radimui plokštumoje, kad tyrėjas galėtų pamatyti ir įvertinti daugiamačių taškų tarpgrupinius panašumus/skirtingumus.

Šiam tikslui pasiekti buvo sprendžiami tokie uždaviniai:

1. analitiškai apžvelgti su darbo tikslu susijusias duomenų tyrybos metodų grupes: vizualizavimo metodų, klasterizavimo metodų ir dirbtinių neuroninių tinklų, o taip pat sukurtus radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginius;
2. išanalizuoti dirbtinių neuroninių tinklų galimybes daugiamačiams duomenims vizualizuoti;
3. optimizuoti radialinių bazinių funkcijų pritaikomumą daugiamačių duomenų matmenų mažinimui remiantis gautų rezultatų vizualia analize;
4. pasiūlyti ir ištirti radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginį (hibridinį tinklą REGM) daugiamačiams duomenims vizualiai tirti, siekiant įvertinti tarpgrupinius panašumus/skirtingumus;
5. pasiūlyti vizualizavimo kokybės kriterijus, kurie padėtų įvertinti gautus vizualizavimo rezultatus;
6. pasiūlyti kriterijus kokybiškai apmokyto REGM tinklo atrankai.

Mokslinis naujumas

Šiame darbe pasiūlytas ir ištirtas naujas hibridinis neuroninis tinklas, kuris savyje integruoja ir radialinių bazinių funkcijų neuroninio tinklo, ir daugiasluoksnio perceptrono, turinčio „butelio kaklelio“ neuroninio tinklo savybes, idėjas. Tai ir yra disertacijos mokslinis naujumas. Toliau šis tinklas bus vadinamas REGM tinklu. Trumpai detalizuosime idėją.

REGM tinklas sudarytas iš dviejų dalių. Pirmojoje dalyje radialinės bazinės funkcijos, kurios atlieka tam tikrą n-matės erdvės $\mathbb{R}^n$ taškų transformavimą į norimo matmens erdvę $\mathbb{R}^k$, $k < n$. Radialinių bazinių funkcijų neuroniniuose tinkluose funkcijų reikšmių apskaičiavimui naudojamą pločio parametrą literatūroje siūloma parinkti pagal tinklo daromą paklaidą. Tačiau šiame darbe pasiūlytame REGM tinkle naudojamos tik radialinės bazinės funkcijos, todėl tinkamam pločio

parametro parinkimui tenka ieškoti kitokių būdų. Šioje disertacijoje pločio parametrą siūloma parinkti pagal objektų išsibarstymą klasteriuose ir vidutinį atstumą tarp tų klasterių centrų. Radialinių bazinių funkcijų pritaikomumas daugiamačių duomenų matmenų mažinimui optimizuotas remiantis gautų rezultatų vizualia analize.

Antroje sudedamojoje REGM tinklo dalyje yra specialios struktūros daugiasluoksnis perceptronas, kurio paskutinis paslėptas sluoksnis yra sudarytas iš nedidelio neuronų skaičiaus (2 arba 3). REGM tinklo paskirtis yra atlikti daugiamačių duomenų projekciją į dvimatę arba trimatę erdvę (projekcija gaunama būtent paskutiniame paslėptame sluoksnyje), kuomet objektus atitinkančius taškus galima stebėti vizualiai. Vizualizuotuose duomenyse atsiskleidžia ir juose esančių klasterių savybės, nes žinios apie klasterių sudėtį, objektus sudarančius klasterius, gaunamos prieš mokant REGM tinklą ir naudojamos to tinklo mokymo metu.

Po REGM tinklo apmokymo vizualiai pateiktos daugiamačių duomenų projekcijos yra įvertinamos šioje disertacijoje užsibrėžtais vizualizavimo kokybės kriterijais. Siekiant galimai geriausio vizualaus duomenų atvaizdavimo, tikslinga REGM tinklą apmokyti keletą kartų ir pasirinkti geriausią projekciją. Spartesniam geriausios duomenų rinkinio projekcijos radimui pagal užsibrėžtus vizualizavimo kokybės kriterijus yra pasiūlyti atrankos kriterijai.

Ginamieji teiginiai

1. Radialinių bazinių funkcijų neuroninio tinklo ir specialios struktūros daugiasluoksnio perceptrono idėjų apjungimas leidžia ieškoti tokios duomenų projekcijos prokštumoje, kad tyrėjas galėtų pamatyti ir įvertinti daugiamačių taškų tarpgrupinius panašumus/skirtingumus.

2. Radialinių bazinių funkcijų pločio parametrą REGM tinklui galima nustatyti pagal objektų išsibarstymą klasteriuose ir vidutinį atstumą tarp tų klasterių centrų.

3. Pasiūlyti trys vizualizavimo kokybės kriterijai įvertina apmokyto REGM tinklo vizualizavimo rezultatus.

4. Jei REGM tinklas apmokomas keletą kartų, geriausios duomenų rinkinio projekcijos pasirinkimą palengvina pasiūlyti du atrankos kriterijai, kuriuos naudojant atranka gali būti automatizuota.

Darbo rezultatų aprobavimas

Tyrimų rezultatai publikuoti 3 periodiniuose recenzuojamuose moksliniuose leidiniuose. Tyrimų rezultatai buvo pristatyti ir aptarti 5 nacionalinėse ir tarptautinėse konferencijose.

Disertacijos struktūra

Disertaciją sudaro 5 skyriai ir literatūros sąrašas. Disertacijos skyriai: Įvadas, Duomenų tyrybos metodai susiję su darbo tikslu ir uždaviniais, REGM tinklas daugiamačiams duomenims vizualizuoti, Eksperimentiniai tyrimai ir Bendrosios išvados. Papildomai disertacijoje pateikta: naudotų žymėjimų ir santrumpų sąrašas. Bendra disertacijos apimtis yra 130 puslapių, kuriuose pateikti 59 paveikslai ir 32 lentelės. Disertacijoje remtasi 101 literatūros šaltiniu.

Apibendrinimas ir bendrosios išvados

Atlikta analitinė hibridinių neuroninių tinklų (įvairūs radialinių bazinių funkcijų ir daugiasluoksnio perceptrono junginiai) apžvalga parodė, kad tokio tipo tinklai konstruojami labai įvairiose srityse ir specifiniams uždaviniams spręsti. Hibridinių tinklų gaunami rezultatai yra tikslesni palyginus su radialinių bazinių funkcijų neuroninių tinklų arba daugiasluoksnių perceptronų gaunamais rezultatais. Konkrečiam uždaviniui spręsti kuriamo hibridinio neuroninio tinklo struktūra pasirenkama pagal atskirų tinklų individualias charakteristikas.

Disertacijoje yra pasiūlytas hibridinis neuroninis tinklas REGM, kuris savyje integruoja ir radialinių bazinių funkcijų neuroninio tinklo, ir daugiasluoksnio perceptrono, turinčio „butelio kaklelio“ neuroninio tinklo savybes, idėjas. Tinklas sudarytas iš dviejų dalių. Pirmoji dalis yra tam tikras daugiamatės erdvės taškų transformavimas į norimo mažesnio matmens erdvę. Antroji dalis yra daugiasluoksnis perceptronas, kurio mažasis sluoksnis (paskutinis paslėptas sluoksnis) sudarytas iš nedidelio neuronų skaičiaus (2 arba 3). REGM tinklo paskirtis yra padėti atskleisti duomenyse esančių klasterių savybes.

REGM tinklas naudojamas vizualiai daugiamačių duomenų analizei, kai atidėjimui plokštumoje arba trimatėje erdvėje taškai gaunami paskutinio paslėpto neuronų sluoksnio išėjimuose į tinklą padavus n-mačių analizuojamų duomenų rinkinį. Šio tinklo ypatybė yra ta, kad gautas vaizdas plokštumoje labiau atspindi bendrą duomenų struktūrą (klasteriai, klasterių tarpusavio artumas, taškų tarpklasterinis panašumas) nei daugiamačių taškų tarpusavio išsidėstymą.

Iš atliktų tyrimų buvo padarytos tokios išvados:

1. REGM tinklas yra nauja efektyvi priemonė daugiamačiams duomenims vizualiai tirti, nes atsiranda galimybė geriau pažinti bendrą duomenų struktūrą. Daugiamačių duomenų klasterizavimo rezultatai gali būti panaudojami ne tik apskaičiuojant radialinių bazinių funkcijų parametrus, bet ir vizualiai pateikiant rezultatus plokštumoje.

2. Jei radialinių bazinių funkcijų (RBF) pločio parametras apskaičiuojamas pagal objektų išsibarstymą klasteriuose ir vidutinį atstumą tarp tų klasterių centrų, tai RBF išėjime gaunamos reikšmės išsibarsto intervale [0; 1], t. y. nesikoncentruoja šio intervalo kraštuose.

3. REGM tinklą apmokius keletą kartų, geriausios duomenų rinkinio projekcijos pasirinkimą palengvina pasiūlyti du atrankos kriterijai, kuriuos naudojant atranka gali būti automatizuota:
   - klasterių išsaugojimas duomenyse kriterijus, kurio reikšmė yra sveikas skaičius ir idealiu atveju lygus 0;
   - išėjimų sluoksnyje gautų taškų išsibarstymo kriterijus, kurio reikšmė yra didžiausias atstumas tarp skirtingiems klasteriams priklausančių taškų.

4. Mažajame sluoksnyje gauta daugiamačių duomenų projekcija yra įvertinama trimis vizualizavimo kokybės kriterijais:
   - taškų išsidėstymas tiesių ar kreivių aplinkoje;
   - taškų „išsibarstymas“ klasteryje (didžiausias atstumas tarp klasterio taškų turi būti didesnis už 0,1);

- riba tarp klasterių (mažiausias atstumas tarp skirtingiems klasteriams priklausančių taškų turi būti didesnis arba lygus 0,05).

5. Disertacijoje visi eksperimentiniai tyrimai su REGM tinklu atlikti naudojant praktinę svarbą turinčius realius duomenų rinkinius, kurių apimtis siekia 4500 objektų. Gautose projekcijose matomi ne tik duomenų rinkinį sudarantys klasteriai, bet ir tarpklasteriniai objektų panašumai/skirtingumai. Skirtinguose klasteriuose esantys, bet panašumų turintys objektai, tyrėjui padeda atkreipti dėmesį į galimus esminius pakitimus objektų savybėse (pavyzdžiui, ankstyvą ligos stadiją arba rūšių panašumus) arba ieškoti priežasčių, dėl kurių atsiranda pakitimai.

6. Hibridinis neuroninis tinklas REGM kokybiškiau apmokomas ir mažajame sluoksnyje gauti vizualizavimo rezultatai atitinka užsibrėžtus vizualizavimo kokybės kriterijus, kai:
   - norimomis tinklo atsako reikšmėmis imami radialinėmis bazinėmis funkcijomis transformuoti klasterių centrai ir radialinių bazinių funkcijų sluoksnyje naudojama eksponentinė, o ne Gausinė funkcija;
   - mažajame sluoksnyje naudojama tiesinė aktyvacijos funkcija, o pirmame paslėptame ir išėjimo sluoksnyje naudojama loginio sigmoido aktyvacijos funkcija;
   - išėjimo sluoksnyje parinktas neuronų skaičius yra lygus pasirinktam klasterių skaičiui.

Trumpos žinios apie autorę

Laura Ringienė gimė 1982 m. rugpjūčio 31 d. Vilniuje. 2001 m. baigė Trakų Vytauto Didžiojo gimnaziją. 2005 m. Vilniaus pedagoginiame universitete įgijo edukologijos bakalauro laipsnį ir mokytojo kvalifikaciją. 2006 m. Vilniaus pedagoginiame universitete įgijo informatikos bakalauro laipsnį ir mokytojo kvalifikaciją. 2008 m. Vilniaus pedagoginiame universitete įgijo informatikos magistro laipsnį. Nuo 2008 iki 2013 buvo Vilniaus universiteto Matematikos ir informatikos instituto doktorantė. Nuo 2005 m. dirba Vilniaus Vytauto Didžiojo gimnazijoje informacinių technologijų mokytoja. Nuo 2008 m. dirba Vilniaus universiteto Matematikos ir informatikos institute inžiniere.

Laura Ringienė

HYBRID NEURAL NETWORK FOR MULTIDIMENSIONAL DATA VISUALIZATION

Summary of Doctoral Dissertation
Technological Sciences,
Informatics Engineering (07 T)

Laura Ringienė

HIBRIDINIS NEURONINIS TINKLAS DAUGIAMAČIAMS DUOMENIMS VIZUALIZUOTI

Daktaro disertacijos santrauka
Technologijos mokslai,
informatikos inžinerija (07 T)