VILNIUS UNIVERSITY

Alma Molytė

INVESTIGATION OF COMBINATIONS OF VECTOR QUANTIZATION METHODS WITH MULTIDIMENSIONAL SCALING

Summary of Doctoral Dissertation
Physical Sciences, Informatics (09 P)

Vilnius, 2011

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2006–2011.

Scientific Supervisors:

Dr Olga Kurasova (Vilnius University, Institute of Mathematics and Informatics, Physical Sciences, Informatics – 09 P) (2007–2011),

Prof Dr Habil Vydūnas Šaltenis (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07 T) (2006–2007).

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

Chairman:

Prof Dr Romas Baronas (Vilnius University, Physical Sciences, Informatics – 09 P).

Members:

Prof Dr Saulius Gudas (Vilnius University, Physical Sciences, Informatics – 09 P),

Prof Dr Habil Rimantas Šeinauskas (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07 T),

Prof Dr Habil Edmundas Kazimieras Zavadskas (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07 T),

Dr Julius Žilinskas (Vilnius University, Physical Sciences, Informatics – 09 P).

Opponents:

Prof Dr Habil Gintautas Dzemyda (Vilnius University, Physical Sciences, Informatics – 09 P),

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 Scientific of Informatics in the auditorium number 203 at the Institute of Mathematics and Informatics of Vilnius University, at 1 p. m. on 29 June 2011.

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

The summary of the doctoral dissertation was distributed on 27th of May 2011.

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

VILNIAUS UNIVERSITETAS

Alma Molytė

VEKTORIŲ KVANTAVIMO METODŲ JUNGIMO SU DAUGIAMATĖMIS SKALĖMIS ANALIZĖ

Daktaro disertacijos santrauka
Fiziniai mokslai, Informatika (09 P)

Vilnius, 2011

Disertacija rengta 2006–2011 metais Matematikos ir informatikos institute.

Moksliniai vadovai:

Dr. Olga Kurasova (Vilniaus universiteto Matematikos ir informatikos institutas, fiziniai mokslai, informatika – 09 P) (2007–2011),

Prof. habil. dr. Vydūnas Šaltenis (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07 T) (2006–2007).

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

Pirmininkas:

prof. dr. Romas Baronas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Nariai:

prof. dr. Saulius Gudas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P),

prof. habil. dr. Rimantas Šeinauskas (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07 T),

prof. dr. habil. Edmundas Kazimieras Zavadskas (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07 T),

dr. Julius Žilinskas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Oponentai:

prof. habil. dr. Gintautas Dzemyda (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P),

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

Disertacija bus ginama viešame Informatikos mokslo krypties tarybos posėdyje 2011 m. birželio mėn. 29 d. 13 val. Vilniaus universiteto, Matematikos ir informatikos instituto 203 auditorijoje.

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

Disertacijos santrauka išsiuntinėta 2011 m. gegužės 27 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje.

Introduction

Relevance of the Problem

Nowadays technologies are able to store and process a large amount of data. However, their perception is a complicated task, especially if the data refer to a complex object or a phenomenon, are defined by many features, and which can be not only numerical, but also logical and textual. Such data are called multidimensional data. Often there is a need to establish and understand the structure of multidimensional data, i. e., their clusters, outliers, similarity and dissimilarity. A set of values of all the features characterize a particular object of the set analyzed. Multidimensional data can be analyzed by various statistical methods. However, if the amount of data is huge, in order to get more knowledge from the data analyzed, various data mining methods (classification, clustering, visualization, etc.) are used.

The area of research is reduction of the number of the data analyzed and mapping the data in a plane (visualization).

A group of methods that enable to discover a new knowledge in the datasets analyzed, is vector quantization methods. The vector quantization is a process when the n-dimensional input vectors are quantized to a limited set of n-dimensional output vectors, the number of which is smaller than that of the input vectors. Commonly these methods are applied in data (sound, image, etc.) compression, but also, they are suitable for clustering and classification.

The target of visualization methods, based on the dimensionality reduction, is to represent the input data in a lower-dimensional space so that certain properties of the dataset were preserved as faithfully as possible. Multidimensional scaling (MDS) refers to a group of methods that are widely used for dimensionality reduction and visualization of multidimensional data. The computational complexity of one iteration of MDS is $O(nm^2)$, where m is the number of data items and n is the number of dimensions. Therefore it is necessary to search ways for acceleration of the computation. The dataset should be reduced so that the new dataset reflected the characteristics of the data analyzed as much as possible.

The results of MDS depend on the initial values of two-dimensional vectors, if the MDS stress is minimized in an iterative way. Various ways of selection of the proper initial values of two-dimensional vectors have been proposed, however, the solution of this task remains a topical problem.

Two main problems are solved here: (1) reduction of the number of data items and their dimensionality, using combinations of the vector quantization methods and multidimensional scaling; (2) investigation of the dependence of MDS results on the ways of selecting the initial values of two-dimensional vectors.

The Aim and Tasks

The aim of the dissertation is to map huge datasets in a lower-dimensional space quickly and precisely, developing a combination of vector quantization and dimensionality reduction methods and investigating the selection of initial values of two-dimensional vectors, which influence the results of visualization.

To achieve the aim, it was necessary to solve the following tasks:

- to analyze the strategies of vector quantization for clustering the datasets;

- to investigate the abilities of combining the vector quantization methods with visualization methods, based on dimensionality reduction;
- to develop new consecutive and integrated combinations of the neural gas and multidimensional scaling and to make their comparison analysis with the combinations of self-organizing map and multidimensional scaling;
- to investigate the ways of selecting the initial values of two-dimensional vectors in the consecutive combination and in the first training block of the integrated combination;
- to investigate the ways of choosing the initial values of two-dimensional vectors in all the training blocks, except the first one, of the integrated combination;
- to analyze the quality of the results of quantization and visualization.

The Objects of Research

The objects of research of the dissertation are vector quantization methods, based on the artificial neural networks, and multidimensional data visualization methods, based on the dimensionality reduction. The following subjects are directly connected with this research object: the measures for evaluating the quality of the projection of multidimensional data into a lower dimensional space; ways of selecting the initial values of two-dimensional vectors.

Scientific Novelty

1. The consecutive and integrated combinations of neural gas and multidimensional scaling have been developed.
2. The ways of selecting the initial values of two-dimensional vectors in the consecutive combination and the first training block of the integrated combination have been proposed and the ways of assigning the initial values of two-dimensional vectors in all the training blocks, except the first one, of the integrated combination have been developed.
3. The dependence of the quantization error on the values of training parameters, the number of epochs, neurons and neuron-winners has been defined experimentally.
4. The fact that combination of the neural gas and multidimensional scaling is more suitable than the combination of the self-organizing map and multidimensional scaling for visualization of the multidimensional data has been experimentally tested and proved.

Practical Significance

The results of investigations, carried out using various real world datasets, have shown that the combination of vector quantization and dimension reduction methods can be widely used to visualize multidimensional data. In the analysis of other real world numerical data, it will be possible to refer to the conclusions, drawn in this dissertation.

Approbation and Publications of the Research

The main results of the dissertation were published in 8 scientific papers: 5 articles in the periodical scientific publications; 3 articles in the proceedings of scientific conferences. The main results of the work have been presented and discussed at 6 national and international conferences.

The Scope of the Scientific Work

The dissertation is written in Lithuanian. It consists of 5 chapters and the list of references. There are 135 pages of the text, 50 figures, 16 tables, and 81 bibliographical sources.

1. Introduction

The relevance of the problems, the scientific novelty of the results and their practical significance as well as the aim and tasks of the work are described in this chapter.

2. Vector Quantization and Visualization Methods

In this chapter the analytic investigation of vector quantization and visualization methods, which are used for multidimensional data visualization, is performed. The vector quantization methods, trained in an unsupervised (neural gas method and self-organizing maps) and supervised (learning vector quantization algorithms) ways, are systematized and analyzed. Vector quantization is used for data compression, missing data correction and clustering. These methods can be combined with visualization methods, known as the projection methods. The main projection methods (multidimensional scaling, principal component analysis) of multidimensional data are analyzed in the dissertation, too. The target of dimensionality reduction methods is to represent the input data in a lower-dimensional space so that certain properties of the dataset were preserved as faithfully as possible.

If we have a dataset $X = \{X_1, X_2, \dots, X_m\}$ in an n-dimensional space, where $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $i = 1, \dots, m$, we desire to get a dataset $Y = \{Y_1, Y_2, \dots, Y_m\}$ in a d-dimensional space, where $Y_i = (y_{i1}, y_{i2}, \dots, y_{id})$, $i = 1, \dots, m$, and $d < n$. If rather a small output dimensionality $d = 2$ or $d = 3$ is chosen, two or three dimensional vectors obtained may be presented in a scatter plot.

The goal of multidimensional scaling (MDS), as one of dimensionality reduction and visualization methods, is to find lower-dimensional data Y_i, $i = 1, \dots, m$, such that the distances between the data in the lower-dimensional space were as close to the original proximities (similarity or dissimilarity) as possible. The MDS error E_{MDS} to be minimized can be written as $E_{MDS} = \sum_{i<j} w_{ij} \left(\delta\left(X_i, X_j\right) - d\left(Y_i, Y_j\right)\right)^2$, where w_{ij} is a weight; $\delta(X_i, X_j)$ is the value of proximity between the n-dimensional data X_i and X_j, $d(Y_i, Y_j)$ is the distance (usually Euclidean) between the two-dimensional data Y_i and Y_j, $d(Y_i, Y_j) = \|Y_i - Y_j\|$. If the proximity is the Euclidean distance, then $\delta\left(X_i, X_j\right) = d\left(X_i, X_j\right)$. We use the SMACOF (Scaling by MAjorization of a COmplicated Function) algorithm for MDS error E_{MDS} minimization, $w_{ij} = 1, \forall\, i, j$. This method guarantees a monotone convergence of the MDS error.

The computational complexity of one iteration of MDS based on SMACOF is $O(nm^2)$. If we analyze a large dataset, MDS is time consuming. Many techniques for reducing the computational time are proposed. Some ways are based on pre-processing: at first, the number m of dataset items is reduced then a smaller dataset is analyzed by MDS. The reduction of m can be done by clustering or vector quantization methods.

When vectors are mapped (visualized), it is necessary to estimate the visualization quality. Three measures were used. The first one is introduced by König. König's

topology preserving measure is based on the assessment of rank order in the n-dimensional and d-dimensional spaces. This measure has two control parameters – numbers of the nearest neighbours: μ and v, $\mu < v$. The neighbourhood is estimated by the Euclidean distances here.

Assume that:

- $X_{ij}, j = 1, \dots, \mu$, are the nearest neighbours of the n-dimensional vector X_i, where the distance X_i and their neighbours satisfy the following inequality $\|X_i - X_{ij_1}\| < \|X_i - X_{ij_2}\|$ with $j_1 < j_2$, here μ is the number of the nearest neighbours;
- $Y_{ij}, j = 1, \dots, v$, are the nearest neighbours of the d-dimensional vector Y_i, v is the number of the nearest neighbours;
- $r_X(i,j)$ is a rank of the jth neighbour X_{ij} of the vector X_i, where the rank means the order number of X_{ij} in the dataset analyzed;
- $r_Y(i,j)$ is a rank of the jth neighbour Y_{ij} of the vector Y_i, corresponding to X_i, where the rank means the order number of Y_{ij} in the dataset analyzed.

König's measure for the ith vector and the jth neighbour is calculated by the formula:

$$E_{\mathrm{KM}}^{ij} = \begin{cases} 3, & \text{if } r_X(i,j) = r_Y(i,j), \\ 2, & \text{if } r_X(i,j) = r_Y(i,l),\ l \in (1, \dots, \mu),\ i \neq l, \\ 1, & \text{if } r_X(i,j) = r_Y(i,t),\ t \in (\mu+1, \dots, v),\ \mu \neq v, \\ 0, & \text{otherwise.} \end{cases}$$

The general König measure E_{KM} is calculated as follows:

$$E_{KM} = \frac{1}{3\mu m} \sum_{i=1}^{\mu} \sum_{j=1}^{m} E_{\mathrm{KM}}^{ij}.$$

The range of E_{KM} is between 0 and 1, where 0 indicates a poor neighbourhood preservation, and 1 indicates a perfect one.

Spearman's rho is calculated by the formula: $\rho_{\mathrm{Sp}}(r'_X, r'_Y) = 1 - \frac{6}{(m')^3 - m'} \sum_{k=1}^{m'} (r'_X(k) - r'_Y(k))^2$, where r'_X and r'_Y are the ranks (order numbers) of pairwise distances calculated for the n-dimensional and d-dimensional data, respectively; $m' = m(m-1)/2$. As usual, $-1 \leq \rho_{Sp} \leq 1$. The best value of Spearman's rho is equal to one.

The third measure is MDS error $\hat{E}_{\mathrm{MDS}} = \sqrt{\frac{\sum_{i<j} (d(X_i, X_j) - d(Y_i, Y_j))^2}{\sum_{i<j} (d(X_i, X_j))^2}}$. This error is used instead of E_{MDS}, because the inclusion of the normalized parameter gives a clear interpretation of the mapping quality that does not depend on the scale in an n-dimensional space.

3. Combination of Vector Quantization and Visualization

The objective of vector quantization for a dataset X is to discover the optimal codebook, containing a predetermined number N of codebook (reference, prototype) vectors $M_i \in R^n$, $i = 1, \dots, N$, which guarantees the minimization of the chosen distortion metric (usually Euclidean) for all the vectors from X. Each codebook vector has an associated index used for referencing. Thus, the aim of quantization is to change

the vectors X_l, $l = 1, \dots, m$, so that the new vectors M_i, $i = 1, \dots, N$, $N < m$, represent the properties of the vectors X_l. Vector quantization is used for data clustering, compression, and missing data correction. In the clustering case, the codebook vectors are representatives of clusters.

The self-organizing map (SOM) is a class of neural networks that are trained in an unsupervised manner using a competitive learning. The neural gas is a biologically inspired adaptive algorithm. The algorithm was named "neural gas" because of the dynamics of the vectors during the adaptation process which distribute themselves like a gas within the data space. The codebook M is an array of vectors. The dimensionality of the vectors is such as that of the analyzed vectors X_l, $l = 1, \dots, m$, i.e., equal to n. The array $M = \{M_1, M_2, \dots, M_N\}$ is one-dimensional in neural gas (NG), $M_i \in R^n$, $i = 1, \dots, N$, N is the number of codebook vectors. The rectangular SOM is a two-dimensional array (grid) of neurons $M = \{M_{ij}, i = 1, \dots, \text{rows}, j = 1, \dots, \text{cols}\}$, where $M_{ij} \in R^n$, rows is the number of rows of the grid, cols is the number of columns of the grid, and the total number of neurons is $N = \text{rows} \times \text{cols}$.

At the beginning of the training algorithms, the initial values are selected: the number N of codebook vectors; the initial values of codebook vector components; the number of training epochs $\hat{e}$ (each analyzed vector is passed to the network $\hat{e}$ times, then the number of training steps $t_{\max} = \hat{e} \times m$).

In NG, the Euclidean distances between the input vector X_l and each codebook vector (neuron) M_i, $i = 1, \dots, N$, are computed. The distances are sorted in an ascending order. A neuron set $W_1, W_2, \dots, W_s$ is obtained, where $W_k \in \{M_1, M_2, \dots, M_N\}$, $k = 1, \dots, N$, and $\|X_l - W_1\| \leq, \dots, \leq \|X_l - W_N\|$. The neuron W_1 is called a winner. The neuron W_k, $k = 1, \dots, N$, is adapted according to the learning rule: $W_k(t+1) = W_k(t) + E(t)h_\lambda(X_l - W_k(t))$, where t is the order number of iterations, $E(t) = E_{\text{g}}(E_{\text{f}}/E_{\text{g}})^{(t/t_{\max})}$, $h_\lambda = e^{-(k-1)/\lambda(t)}$, $\lambda(t) = \lambda_{\text{g}}(\lambda_{\text{f}}/\lambda_{\text{g}})^{(t/t_{\max})}$. The values of the parameters λ_{g}, λ_{f}, E_{g}, E_{f} are predetermined.

In SOM, the Euclidean distances from the input vector X_l to each codebook vector M_{ij}, $i = 1, \dots, \text{rows}, j = 1, \dots, \text{cols}\}$, are computed as well. The vector (neuron) $\widehat{M}_c$ with the minimal Euclidean distance to X_l is designated as a winner, where c is a pair of indices, i.e., $c = \arg\min_{i,j}\{\|X_l - M_{ij}\|\}$. The neuron M_{ij} is adapted according to the learning rule: $M_{ij}(t+1) = M_{ij}(t) + h_{ij}^c(t)(X_l - M_{ij}(t))$, where t is the order number of iterations, h_{ij}^c is a neighbourhood function, $h_{ij}^c(t) \to 0$, as $t \to \infty$. There are a lot of variants of h_{ij}^c. We use $h_{ij}^c = \alpha/(\alpha\eta_{ij}^c + 1)$, $\alpha = \max((\hat{e} + 1 - e')/\hat{e},\ 0.01)$; η_{ij}^c is the neighbourhood order in the grid between the neurons M_{ij} and $\widehat{M}_c$; $\hat{e}$ is the number of training epochs, e' is the order number of the current epoch ($e' \in \{1, \dots, \hat{e}\}$). The vector M_{ij} is recomputed, if $\eta_{ij}^c \leq \max[\alpha\ \max(\text{rows}, \text{cols}),\ 1]$. For generality, the notation M_i is used instead of M_{ij} below.

Then the networks are trained, the quantization error E_{QE} is computed by the formula $E_{\text{QE}} = \frac{1}{m}\sum_{l=1}^{m}\|X_l - \widehat{M}_{c(l)}\|$, where $\widehat{M}_{c(l)}$ is a winner for the vector X_l, $\widehat{M}_{c(l)} = W_1$ in the neural gas method.

After training the NG or SOM network, each input vector X_i, $i = 1, \dots, m$, from X is related to the nearest neuron, called a neuron-winner. Some neurons may remain

unrelated with any vector of the set X, but there may occur neurons related with some input vectors. So, the neuron-winners represent some input vectors, and the number r of neuron-winners is smaller than that of input vectors ($r < m$). Thus, the number m of data items is reduced. A smaller dataset can be used by MDS and the time is saved.

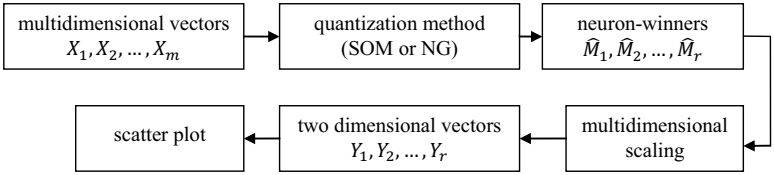

Fig. 1. *The scheme of visualization of neuron-winners (consecutive combination)*

So, the reason for using a consecutive combination (Fig. 1) is a desire to decrease the computation time without losing the quality of mapping (visualization).

Another reason is based on improving the SOM visualization. As it is known, the SOM itself has a visual presentation, e. g., u-matrix representation. However, the SOM table does not answer the question, how much the vectors of the neighbouring cells are close in the n-dimensional space. It is reasonable to apply the distance-preserving method, such as MDS, to an additional mapping of the neuron-winners in SOM. A question arises: when the usage of MDS only is purposeful, and when its combination with vector quantization.

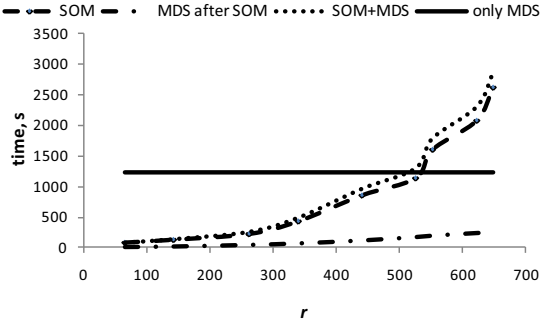

Fig. 2. *The computational time of MDS only and its combination with SOM*

The computing time of MDS only, when all the items of the ellipsoidal dataset ($m = 1338,\ n = 100$) have been analyzed is presented in Fig. 2 (black solid line). The SOM learning has been repeated for several times with various numbers N of neurons. Various numbers r of neuron-winners have been obtained. The dependence of the SOM learning time on the number r of neuron-winners (dashed curve), as well as of MDS on the number r of neuron-winners, when only they are analyzed by MDS (dashed with point curve), and the total time of the SOM and MDS combination (dotted curve) are presented in Fig. 2. We see that if the number r of neuron-winners is smaller than 500, it

is worth to using the combination in order to save the computational time as compared with MDS only. If NG is used instead of SOM, the similar results are obtained.

The visualization results of the ellipsoidal dataset when all data items ($m = 1338$) are mapped by MDS and only 262 neuron-winners ($r = 262$) of SOM are mapped by MDS are presented in Fig. 3. We see that reduction of the number of data items does not aggravate the quality of visualization, while the computing time is saved essentially.

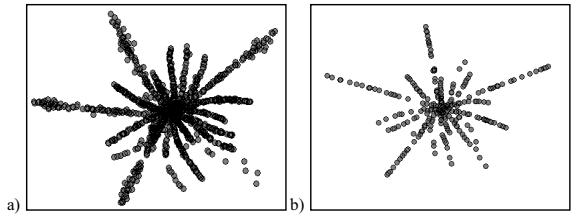

Fig. 3. *Mapping of an ellipsoidal dataset: a) all data items are mapped by MDS; b) only 262 neuron-winners of SOM are mapped by MDS*

Note that, if the MDS error E_{MDS} is minimized in an iterative way, it is important to select the proper initial values of d-dimensional vectors $Y_1, Y_2, \dots, Y_m$ (in our case, $d = 2$). The dependence of the MDS results on the initial values of these vectors remains a topical problem. We have proposed and investigated the integrated combination of SOM and MDS as a new way of initialization of two-dimensional vectors. We suggest to use NG instead of SOM.

The idea of the integrated combination is as follows: n-dimensional vectors $X_1, X_2, \dots, X_m$ are analyzed by using the MDS method, taking into account the process of SOM or NG training. Thus, the integrated combination consists of two parts: (1) SOM or NG training; (2) computing two-dimensional points, corresponding to the neuron-winners of SOM or NG, by the MDS method. These two parts are performed alternately.

At first, some notation and definitions are introduced:

- Let the training set consist of n-dimensional vectors $X_1, X_2, \dots, X_m$, $(X_i = (x_{i1}, x_{i2}, \dots, x_{in}),\ i = 1, \dots, m)$. We need to get two-dimensional vectors, called projections, $Y_1, Y_2, \dots, Y_m$, $(Y_i = (y_{i1}, y_{i2}),\ i = 1, \dots, m)$.
- The neural network (SOM or NG) is trained using $\hat{e}$ training epochs.
- All the $\hat{e}$ epochs are divided into equal training parts – blocks. Before starting the training of the neural network, we choose the number of blocks γ into which the training process will be divided. Each block contains ν training epochs ($\hat{e} = \nu\gamma$). Denote by q a block of the training process consisting of ν epochs ($q = 1, \dots, \gamma$).
- Denote neuron-winners, obtained by the qth block of the training process, as $M_1^{(q)}, M_2^{(q)}, \dots, M_{r_q}^{(q)}$ and two-dimensional projections of these neuron-winners, calculated by the MDS method, as $Y_1^{(q)}, Y_2^{(q)}, \dots, Y_{r_q}^{(q)}$ $(Y_i^{(q)} = (y_{i1}^{(q)}, y_{i2}^{(q)})$, $i = 1, .., r_q)$. Note that the number of neuron-winners r_q will be smaller than or equal to m.

We suggest the following way of integrating the SOM or NG and MDS methods (Fig. 5):

Step 1: network training begins ($q = 1$). After the first ν training epochs, the training is stopped temporally. The neuron-winners $M_1^{(1)}, M_2^{(1)}, \dots, M_{r_1}^{(1)}$, obtained after the first block ($q = 1$) of the training process, are analyzed by MDS. The initial coordinates of two-dimensional vectors ($Y_i^{(0)} = (y_{i1}^{(0)}, y_{i2}^{(0)})$, $i = 1, \dots, r_1$) must be set for MDS. There are some possible ways. The initial coordinates ($y_{i1}^{(0)}, y_{i2}^{(0)}$) can be set:

1. At random in the interval $(0; 1)$.
2. On a line: $y_{i1}^{(0)} = i + 1/3$, $y_{i2}^{(0)} = i + 2/3$.
3. According to two largest principal components (PCs).
4. According to the components whose variances are the largest ones.

After MDS has been performed, the two-dimensional projections $Y_1^{(1)}, Y_2^{(1)}, \dots, Y_{r_1}^{(1)}$ of neuron-winners are obtained.

Steps from 2 to γ: network training is continued ($q = 2, \dots, \gamma$). The neuron-winners obtained after each qth block of the training process are analyzed by using MDS. The initial coordinates of two-dimensional vectors $Y_1^{(q)}, Y_2^{(q)}, \dots, Y_{r_q}^{(q)}$ are selected for MDS taking into account the result of the $(q-1)$ block. Note that $r_q \neq r_{q-1}$ in general. The way of selecting the initial coordinates is presented below. We must determine the initial coordinates of each two-dimensional vector $Y_i^{(q)}$ correspondent to the neuron-winner $M_i^{(q)}$, $i = 1, .., r_q$. The sequence of steps is as follows:

- Determine vectors from $\{X_1, X_2, \dots, X_m\}$ that are related with $M_i^{(q)}$. Note that some vectors from $\{X_1, X_2, \dots, X_m\}$ can be related with $M_i^{(q)}$. Denote these vectors by $X_{i1}, X_{i2}, \dots$ ($X_{i1}, X_{i2}, \dots \in \{X_1, X_2, \dots, X_m\}$).
- Determine neuron-winners of the $(q-1)$ block that were related with $X_{i1}, X_{i2}, \dots$ Denote these neuron-winners by $M_{j_1}^{(q-1)}, M_{j_2}^{(q-1)}, \dots$ ($M_{j_1}^{(q-1)}, M_{j_2}^{(q-1)}, \dots \in \{M_1^{(q-1)}, M_2^{(q-1)}, \dots, M_{r_{q-1}}^{(q-1)}\}$), and their two-dimensional projections, obtained as a result of MDS, by $Y_{j_1}^{(q-1)}, Y_{j_2}^{(q-1)}, \dots$ ($Y_{j_1}^{(q-1)}, Y_{j_2}^{(q-1)}, \dots \in \{Y_1^{(q-1)}, Y_2^{(q-1)}, \dots, Y_{r_{q-1}}^{(q-1)}\}$).
- There are two possible ways of assignment (Fig.6):

  by proportion: the initial coordinates of $Y_i^{(q)}$ are set to be equal to the mean value of the set of vectors $\{Y_{j_1}^{(q-1)}, Y_{j_2}^{(q-1)}, \dots\}$. In Fig. 4 (top), two points $Y_{j_1}^{(q-1)}$ and $Y_{j_2}^{(q-1)}$ are coincident, the point $Y_i^{(q)} = 1/3\,(Y_{j_1}^{(q-1)} + Y_{j_2}^{(q-1)} + Y_{j_3}^{(q-1)})$ is closer to the points $Y_{j_1}^{(q-1)}$ than to $Y_{j_3}^{(q-1)}$.

  by midpoint: since the coincident vectors can be between the vectors $Y_{j_1}^{(q-1)}, Y_{j_2}^{(q-1)}, \dots$ the initial coordinates of $Y_i^{(q)}$ are set to be equal to the mean value of the set of only the non-coincident points $Y_{j_1}^{(q-1)}, Y_{j_2}^{(q-1)}, \dots$ In Fig. 4 (bottom), $Y_i^{(q)} = 1/2\,(Y_{j_1}^{(q-1)} + Y_{j_3}^{(q-1)})$.

After the assignment, the two-dimensional vectors $Y_1^{(q)}, Y_2^{(q)}, \dots, Y_{r_q}^{(q)}$ ($Y_i^{(q)} = (y_{i1}^{(q)}, y_{i2}^{(q)})$, $i = 1, \dots, r_q$) of the neuron-winners are calculated using MDS.

The training of the neural network is continued until $q = \gamma$. After the γth block, we get two-dimensional projections $Y_1^{(\gamma)}, Y_2^{(\gamma)}, \dots, Y_{r_\gamma}^{(\gamma)}$ of the n-dimensional neuron-winners $M_1^{(\gamma)}, M_2^{(\gamma)}, \dots, M_{r_\gamma}^{(\gamma)}$ that are uniquely related with the vectors $X_1, X_2, \dots, X_m$. The two-dimensional vectors $Y_1^{(\gamma)}, Y_2^{(\gamma)}, \dots, Y_{r_\gamma}^{(\gamma)}$ obtained can be presented on a scatter plot.

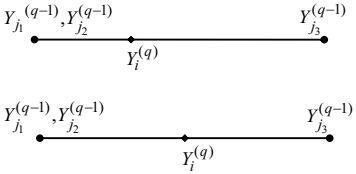

Fig. 4. *Two ways of assignment: by proportion (top), by midpoint (bottom)*

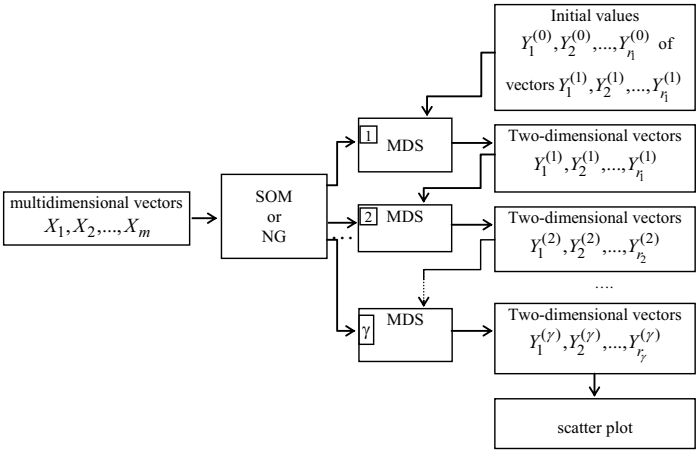

Fig. 5. *The scheme of the integrated combination of SOM or NG and multidimensional scaling*

4. Experimental Investigations

In this chapter, the results of experimental investigations of two vector quantization methods (neural gas and self-organizing maps) and their combinations with multidimensional scaling are presented. The iris [150; 4], hepta [212; 3], auto MPG [392; 7], target [770; 2], chainlink [1000; 3], rand_clust10 [100; 10], rand_clust5 [100; 5] and rand_data1500 [1500;5] datasets are used in the experimental investigations.

The numbers of the neural gas and the self-organizing maps are investigated. It is of interest to investigate by which method (NG or SOM) more neurons become winners.

The ratios between the number of neuron-winners and all the neurons of NG and SOM are presented in Fig. 6. It is shown that the ratios of NG are larger than that of SOM: about 80 % of the NG neurons become winners. If the numbers of neurons are large, only about 50 % of the SOM neurons become winners. The investigation shows that SOM is more useful than the neural gas for solving clustering problems.

The quantization error E_{QE} is calculated to estimate the quality of quantization. The quantization error shows the difference between the analyzed vectors $X_1, X_2, \ldots, X_m$ and the quantized vectors (neuron-winners) $\widehat{M}_1, \widehat{M}_2, \ldots, \widehat{M}_r$, where r is the number of neuron-winners.

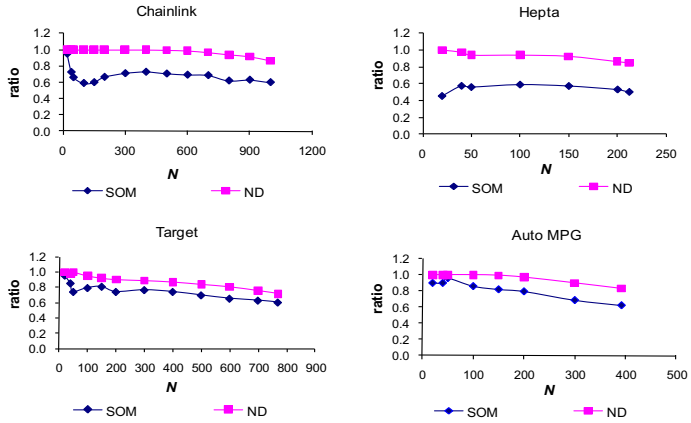

Fig. 6. *The ratio of neurons and neuron-winners in NG and SOM*

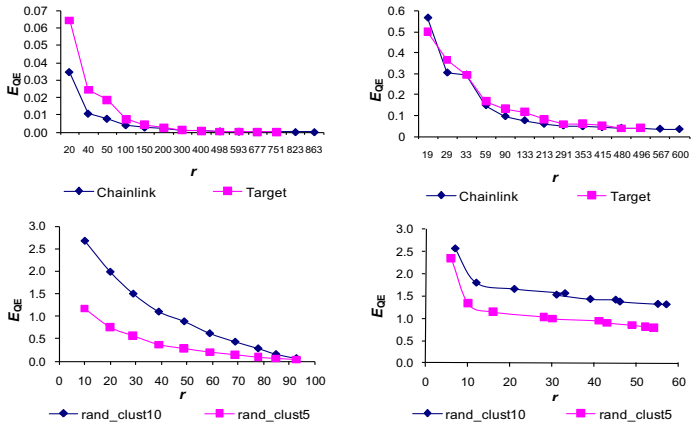

Fig. 7. *Dependence of quantization error on number of neuron-winners, obtained by NG (left) and SOM (right)*

The dependence of the quantization error on the number of neuron-winners is presented in Fig. 7. The quantization error decreases, if the number of neuron-winners is increasing. As we see in Fig. 7, the quantization errors of NG are significantly smaller than that of SOM when the number of neuron-winners is approximately equal. It means that the neural gas is more suitable for vector quantization.

König's topology preservation measure E_{KM} and Spearman's rho ρ_{Sp} are calculated to estimate the visualization quality. The number N of codebook vectors is selected so that the number of neuron-winners were equal to 100, 200, and 300 for the chainlink and auto MPG, to 50, 100, and 150 for the iris, to 50, 100, and 200 for the hepta, and to 50, 80, and 100 for the rand_clust10 datasets.

Since the results of SOM and NG depend on the initial values of codebook vectors, 40 experiments have been carried out for each input vector set with different initial values of codebook vectors. The values of the measures are calculated and averaged. The confidence intervals of the averages are also calculated (a probability is equal to 0.95).

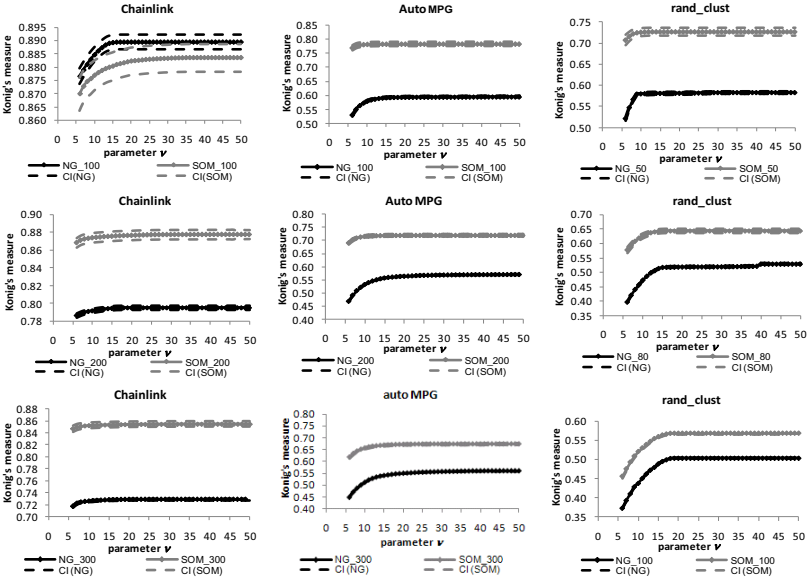

Fig. 8. *Dependence of values of König's measure on parameter* ν

When calculating König's topology preserving measure E_{KM}, it is necessary to select values of two parameters μ and ν. The parameter μ indicates a narrow round of neighbours, and the parameter ν indicates a wide round. In the experiments, $\mu = 4$, and ν is varying from 6 to 50. The averaged values of E_{KM} and the confidence intervals (CI) of the averages are presented in Fig. 8. We see that E_{KM} is larger, if the neuron-winners obtained by SOM are mapped in all the cases, except the chainlink dataset, where the number of neuron-winners is equal to 100. We conclude that the topology is preserved precisely when the vector-winners obtained by SOM are mapped by MDS. In an exceptional case, the confidence intervals are wide, they are overlapping, and therefore

the results obtained are unreliable. When the number of neuron-winners is increasing, the confidence intervals are narrowing for all datasets. Naturally, for small values of the parameter ν, the values of E_{KM} are lower than that for higher ν, however starting from a certain value of ν, the values of E_{KM} do not change at all or change but slightly.

The averaged values of Spearman's rho ρ_{Sp} and the confidence intervals (CI) of the averages are presented in Fig. 9. The values of Spearman's rho are higher, if the neuron-winners are obtained by NG for the chainlink and hepta datasets, and by SOM for the auto MPG and iris datasets. The values of Spearman's rho are large enough (in many cases, $\rho_{\mathrm{Sp}} > 0.9$), which means that the mapping results are good in the sense of distance preserving, when passing from the n-dimensional space to a two-dimensional one. It is difficult to draw a conclusion on the mapping quality of the rand_clust10 dataset, because the values of Spearman's rho are varying, and the confidence intervals are wide and overlapping. The investigation shows that both the NG and SOM methods are suitable for a combination with MDS.

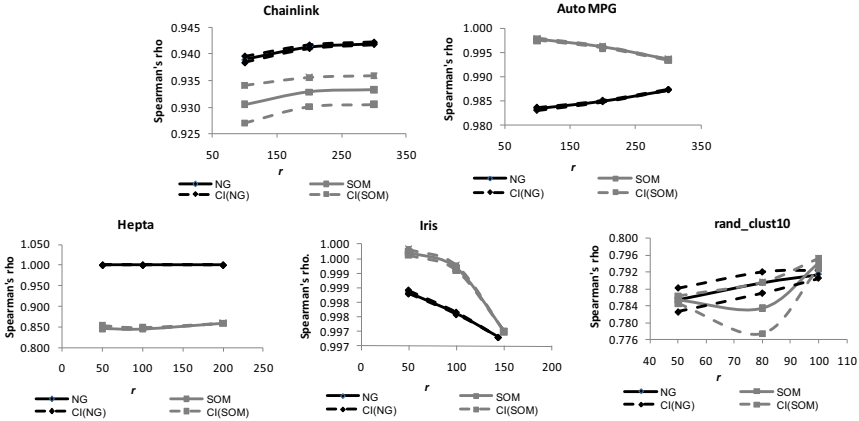

Fig. 9. *Dependence of values of Spearman's rho on the number of neuron-winners*

Two-dimensional vectors may be presented in a scatter plot. The mapping images of chainlink dataset are presented in Fig. 10. The numbers near the points indicate the order numbers of classes to which the points belong. Fig. 10 shows how the mapping images change when the number of neuron-winners is growing. The data structure is visible even when the number r of neuron-winners, obtained by NG, is small enough. If the number r of neuron-winners, obtained by SOM, is larger, the data structure is visible, as well.

In Fig. 11, the neuron-winners of the iris dataset, obtained by NG and SOM, are visualized by MDS. The points, corresponding to the items of the first species (Setosa), are marked by filled rhombi, the points, corresponding to the second species (Versicolor), are marked by filled squares and the points, corresponding to the third species (Virginica), are marked by filled circles. The points, corresponding to the neurons, that are the winners for both the second and third species, are marked by boxed circles. The quantization error of SOM is much larger ($E_{\mathrm{QE}} = 0.3222$) than that of NG

($E_{\mathrm{QE}} = 0.0379$). It means that the neuron-winners (quantized vectors) do not approximate the data by SOM precisely enough. We see that the points obtained by SOM are clustered very much, but the points obtained by NG are dispersed. The data structure is revealed better by NG.

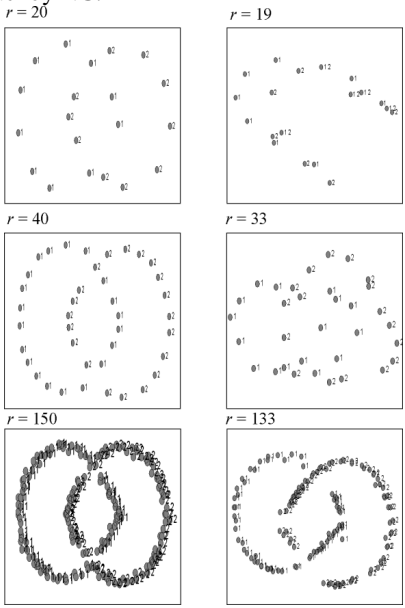

Fig. 10. *Mapping images of chainlink data obtained by NG (left) and SOM (right)*

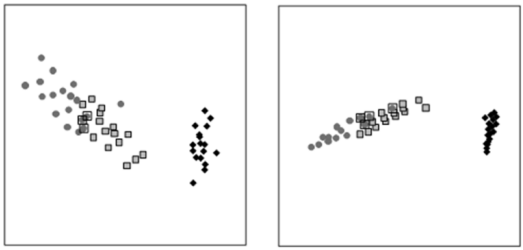

Fig. 11. *Mapping images of iris data obtained by NG (left)* ($E_{QE} = 0.0379$) *and SOM (right)* ($E_{QE} = 0.3222$)

Some experiments have been done in order to ascertain which vector quantization method (SOM or NG) is more suitable to use in the combination with MDS and which initialization way of two-dimensional points is most suitable in the consecutive combination of SOM or NG and MDS, as well as in the first block of the integrated combination (when the points are generated at random, on a line, according to two principal components (PCs), according to the components with the largest variances);

which way of assignment in the integrated combination is the most suitable one (by midpoint or by proportion).

The results of experimental investigation of some datasets are presented here: iris ($m = 150$, $n = 4$), hepta ($m = 212$, $n = 3$), and rand data ($m = 1500$, $n = 5$) (here each component is generated at random in the interval (0;1)). SOM and NG are trained during 200 epochs ($\hat{e} = 200$). The training process is divided into $\gamma = 2, 4, 8, 10, 25$ blocks in the integrated combination and $\nu = 100, 50, 25, 20, 8$, respectively. 100 iterations are performed in MDS. The values of the MDS error $\hat{E}_{\mathrm{MDS}}$ subject to the initialization and assignment ways for three datasets are presented in Tables 1–3. When choosing a random initialization, ten experiments are done for each dataset and the averaged values are presented in Tables 1–3 and Fig. 12. The smallest values are in italics and the most frequent values are in bold. The number N of neurons is set such that the same or a similar number r of neuron-winners were obtained by both vector quantization methods with a view to compare the results obtained in the sense of the MDS error $\hat{E}_{\mathrm{MDS}}$.

When comparing the results, obtained by the consecutive and integrated combinations, smaller values of the MDS error are obtained by the integrated combination in many cases. Thus, the integrated combination is superior to the consecutive one. It is quite evident, if the points are initiated on a line or at random (Fig. 12). The values of the MDS error, obtained by the consecutive combination and the smallest values of the error, obtained by the integrated combination, are presented in Fig. 12.

In most cases, the MDS error is slightly larger, if NG is used instead of SOM in combinations. However, the quantization error E_{QE} is considerably smaller, therefore NG is more suitable in the combinations.

Table 1. *Values of the MDS error subject to the initialization and assignment ways for the iris dataset*

a) SOM ($E_{\mathrm{QE}} = 0.2225$, $r = 93$)

consecutive			at random		on a line		by PCs		by variances	
			0.0363		0.0366		0.0276		0.0265	
integrated	γ	ν	midpoint	proportion	midpoint	proportion	midpoint	proportion	midpoint	proportion
	2	100	0.0385	0.0386	0.0484	0.0484	0.0395	0.0436	0.0438	0.0438
	4	50	0.0371	0.0373	***0.0265***	0.0271	0.0382	0.0269	0.0382	0.0382
	8	25	0.0335	0.0296	***0.0265***	***0.0265***	***0.0265***	***0.0265***	0.0347	***0.0265***
	10	20	0.0281	***0.0265***	0.0347	***0.0265***	***0.0265***	***0.0265***	***0.0265***	***0.0265***
	25	8	0.0298	0.0290	0.0347	***0.0265***	0.0347	***0.0265***	0.0347	***0.0265***

b) NG ($E_{\mathrm{QE}} = 0.0988$, $r = 94$)

consecutive			at random		on a line		by PCs		by variances	
			0.0489		0.0642		0.0335		0.0358	
integrated	γ	ν	midpoint	proportion	midpoint	proportion	midpoint	proportion	midpoint	proportion
	2	100	0.0451	0.0452	0.0381	0.0561	***0.0335***	***0.0335***	***0.0335***	***0.0335***
	4	50	0.0399	0.0417	***0.0335***	***0.0335***	***0.0335***	***0.0335***	***0.0335***	***0.0335***
	8	25	0.0366	0.0363	***0.0335***	***0.0335***	***0.0335***	***0.0335***	***0.0335***	***0.0335***
	10	20	0.0392	0.0384	***0.0335***	***0.0335***	***0.0335***	0.0349	0.0349	0.0349
	25	8	0.0369	0.0388	0.0506	***0.0335***	***0.0335***	***0.0335***	***0.0335***	***0.0335***

When the number γ of blocks of the integrated combination is increased, the MDS error is rather fluctuating, however it is no larger than that obtained by the consecutive combination.

The smallest value of the MDS error for the iris dataset is obtained, if the initial values of two-dimensional points are set by variances, when SOM is used in the consecutive combination, $\widehat{E}_{\mathrm{MDS}} = 0.0265$, and by principal components, when NG is used $\widehat{E}_{\mathrm{MDS}} = 0.0335$. However, the same minimal value of the MDS error is obtained by the integrated combination, when other initialization ways are used.

Table 2. *Values of the MDS error subject to the initialization and assignment ways for the hepta dataset*

a) SOM ($E_{\mathrm{QE}} = 0.3115, r = 86$)

consecutive			at random		on a line		by PCs		by variances	
			0.2182		0.2270		0.2042		0.2042	
integrated	γ	ν	midpoint	proportion	midpoint	proportion	midpoint	proportion	midpoint	proportion
	2	100	0.2004	0.2066	***0.1994***	***0.1994***	***0.1994***	***0.1994***	***0.1994***	***0.1994***
	4	50	0.2078	**0**.2345	***0.1994***	***0.1994***	***0.1994***	0.2042	0.2270	0.2487
	8	25	***0.1994***	0.2109	***0.1994***	0.2270	***0.1994***	***0.1994***	***0.1994***	0.2270
	10	20	***0.1994***	0.2051	***0.1994***	0.2042	***0.1994***	***0.1994***	***0.1994***	0.2042
	25	8	***0.1994***	0.2081	***0.1994***	***0.1994***	***0.1994***	***0.1994***	***0.1994***	***0.1994***

b) NG ($E_{\mathrm{QE}} = 0.1765, r = 94$)

consecutive			at random		on a line		by PCs		by variances	
			0.2053		0.2115		0.1964		0.1964	
integrated	γ	ν	midpoint	proportion	midpoint	proportion	midpoint	proportion	midpoint	proportion
	2	100	0.1877	0.1877	0.2043	0.2043	**0.1964**	**0.1964**	0.2043	0.2043
	4	50	0.2084	0.2084	0.2322	0.2322	0.2043	0.2043	0.2056	0.2056
	8	25	0.2194	0.2194	**0.1964**	**0.1964**	**0.1964**	**0.1964**	**0.1964**	**0.1964**
	10	20	0.2008	0.2052	**0.1964**	**0.1964**	**0.1964**	**0.1964**	**0.1964**	**0.1964**
	25	8	0.2115	0.2031	0.2115	**0.1964**	0.2115	**0.1964**	0.2115	**0.1964**

Table 3. *Values of the MDS error subject to the initialization and assignment ways for the rand_data1500 dataset*

a) SOM ($E_{\mathrm{QE}} = 0.2139, r = 394$)

consecutive			at random		on a line		by PCs		by variances	
			0.3223		0.3189		0.3153		0.3140	
integrated	γ	ν	midpoint	proportion	midpoint	proportion	midpoint	proportion	midpoint	proportion
	2	100	0.3244	0.3247	0.3252	0.3237	0.3241	0.3239	0.3241	0.3216
	4	50	0.3217	0.3225	0.3217	0.3220	0.3217	0.3220	0.3218	0.3229
	8	25	0.3176	0.3200	0.3178	0.3148	0.3176	*0.3142*	0.3177	0.3206
	10	20	0.3157	0.3155	0.3164	0.3162	0.3164	0.3164	0.3164	0.3167
	25	8	0.3159	**0.3161**	0.3162	**0.3161**	0.3160	**0.3161**	0.3162	**0.3161**

b) NG ($E_{\mathrm{QE}} = 0.1380, r = 400$)

consecutive			at random		on a line		by PCs		by variances	
			0.3202		0.3223		0.3119		0.3103	
integrated	γ	ν	midpoint	proportion	midpoint	proportion	midpoint	proportion	midpoint	proportion
	2	100	0.3192	0.3143	0.3179	0.3179	0.3125	0.3123	0.3140	0.3116
	4	50	0.3168	0.3159	0.3160	0.3160	0.3183	0.3187	**0.3115**	0.3140
	8	25	0.3129	0.3122	0.3132	0.3157	**0.3115**	**0.3115**	*0.3103*	0.3115
	10	20	0.3124	0.3131	0.3116	0.3223	0.3116	0.3119	0.3115	*0.3103*
	25	8	**0.3115**	**0.3115**	**0.3115**	0.3220	**0.3115**	**0.3115**	**0.3115**	**0.3115**

The smallest value of the MDS error $\hat{E}_{\mathrm{MDS}} = 0.1994$ for the hepta dataset is obtained by the integrated SOM and MDS combination independent of the initialization way. When NG is used, the most frequent value $\hat{E}_{\mathrm{MDS}} = 0.1964$ is obtained by the consecutive combination, if the initial values are set by variances or principal components. The same value is obtained by the integrated combination, if the initial values are set on a line. If the random initialization is used, the smallest value $\hat{E}_{\mathrm{MDS}} = 0.1877$ is obtained by the integrated combination, $\gamma = 2$.

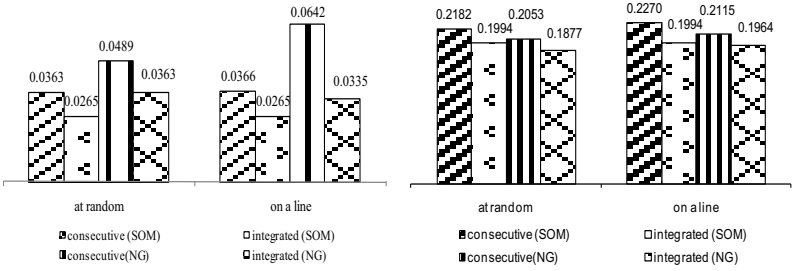

Fig. 12. *Values of the MDS error, obtained by the consecutive and integrated combinations, for the iris dataset (left) and the hepta dataset (right)*

There is no value of the MDS error that could be minimal and repeated for the rand_data1500 dataset in contrast to the iris and hepta ones. However, the tendency of error decline is shown in the integrated combination, when the number γ of blocks is increased.

When two ways of assignment (by midpoint and proportion) in the integrated combination are compared, no great difference was noticed.

General Conclusions

The research results have shown new capabilities of the combination of vector quantization methods – self-organizing maps and neural gas – and multidimensional scaling. The results of the experimental research allow us to draw the following conclusions:

1. Approximately 80 % of neurons become neuron-winners by the neural gas method and approximately 50 % neurons become neuron-winners by the SOM method, therefore SOM is more useful for data clustering. However, the quantization errors, obtained by the neural gas method, are smaller than the errors, obtained by the SOM method, the number of neuron-winners being approximately equal. Thus, the neural gas method is more suitable for vector quantization as well as for the usage in the combination of multidimensional scaling.
2. In the combination of multidimensional scaling and SOM, the neighbourhood relations are preserved more precisely in the sense of König's measure than in the case, where the neural gas method is used. Both quantization methods are equivalent in the sense of Spearman's rho. The MDS error is smaller, when SOM is used instead of the neural gas method in the combination, for data, dimension of which $n = 5,\ 7,\ 10$.

3. When the proposed integrated combination of the neural gas and multidimensional scaling is applied, the data structure is visible, if the training blocks $q < 1/3\,\gamma$, and using SOM instead of the neural gas method in the integrated combination, the number of training blocks must be $q > 2/3\,\gamma$, where γ is the number of all the training blocks.
4. The proposed assignment of the initial values of two-dimensional vectors by midpoint in the integrated combination, except the first training block, can be used as an alternative of the assignment by proportion, because no essential difference is observed in the results obtained.
5. The MDS error, obtained by the consecutive combination, is the smallest one, when the initial values of two-dimensional vectors are selected by two principal components or by the components with the largest variances. Sometimes it is possible to reduce the error even more using the integrated combination.

List of the author's publications on the subject of the dissertation

Articles in the reviewed scientific periodical publications:

1. Kurasova, O.; Molytė A., 2008. Selection of the Number of Neurons for Vectors Quantization Methods, *Lithuanian Mathematical Journal*, T. 48/49, 354–359. ISSN 0132-2818 (in Lithuanian).
2. Molytė, A.; Kurasova O., 2009. Combination of Vector Quantization and Multidimensional Scaling Methods for Visualization of Multidimensional Data, *Information Sciences*, Vilnius, Vilnius University, T. 50, 340–346. ISSN 1392-0561 (in Lithuanian).
3. Kurasova, O.; Molytė, A., 2009. Combination of Vector Quantization and Visualization, In: P. Perner (Ed.,) Machine Learning and Data Mining in Pattern Recognition - MLDM 2009, *Lecture Notes in Artificial Intelligence*, Springer Verlag, Heidelberg, Vol. 5632, 29–43. ISSN 0302-9743, ISBN 978-3-642-03069-7.
4. Kurasova, O.; Molytė, A., 2011. Quality of Quantization and Visualization of Vectors Obtained by Neural Gas and Self-organizing Map, *Informatica*, Vilnius University, Vol. 22 (1), 115–134. ISSN 0868-4952. (ISI Web of Science, Impact Factor 2009: 1.040)
5. Kurasova, O.; Molytė, A., 2011. Integration of the Self-organizing Map and Neural Gas With Multidimensional Scaling, *Information Technology and Control*, Vol.40 (1). ISSN 1392-124X. (ISI Web of Science, Impact Factor 2009: 0.495)

Articles in other editions:

1. Molytė, A.; Kurasova, O., 2008. Investigation of the Vector Quantization Method "Neural-Gas", *The 11th Lithuanian Young Scientists Conference "Science – The future of Lithuania" Thematic Conference "Informatics"a collection of articles for 2008*, Vilnius, VGTU, 198–205. ISBN 978-9955-28-302-7 (in Lithuanian)..
2. Kurasova, O.; Molytė, A., 2009. Investigation of the Quality of Mapping Vectors Obtained by Quantization Methods, *Proceedings of the XIII International Conference on Applied Stochastic Models and Data Analysis (ASMDA-2009)*, Selected papers (L. Sakalauskas, C. Skiadas, E. K. Zavadskas (eds.)). Vilnius, Technika, 269–273. ISBN 978-9955-28-463-5.

3. Kurasova, O.; Molytė, A., 2009. Integrated Visualization of Vector Quantization by Multidimensional Scaling. *Abstract Proceedings of Workshop on Computer Graphics, Vision and Mathematics (GraVisMa 2009)* (D. Hildenbrand. V. Skala (eds.)), University of West Bohemia, 22. ISBN 978-80-86943-92-3.

Short description about the author

Alma Molytė received a Bachelor's degree in mathematics from the Vilnius Pedagogical University in 1999 and Master's degree in informatics in 2001. 2006–2011 – PhD studies at the Institute of Mathematics and Informatics, System Analysis Department. She is a member of the Lithuanian Computer Society.

VEKTORIŲ KVANTAVIMO METODŲ JUNGIMO SU DAUGIAMATĖMIS SKALĖMIS ANALIZĖ

Tyrimų sritis ir problemos aktualumas

Dabartinėmis technologijomis galima gauti ir saugoti didelius duomenų kiekius, tačiau jų suvokimas gana sudėtingas uždavinys, ypač kai duomenys nurodo sudėtingą objektą ar reiškinį, kuris aprašytas daugeliu parametrų, ir kurie gali būti ne tik skaitiniai, bet ir loginiai bei tekstiniai. Tokie duomenys vadinami daugiamačiais duomenimis. Dažnai iškyla būtinybė nustatyti ir giliau pažinti šių daugiamačių duomenų struktūrą, t. y. susidariusius klasterius, itin išsiskiriančius objektus, objektų tarpusavio panašumą ir skirtingumą. Visų parametrų reikšmių junginys charakterizuoja vieną analizuojamos aibės konkretų objektą. Daugiamačiai duomenys gali būti analizuojami įvairiais statistikos metodais, tačiau kai duomenų kiekis yra didelis, dažnai jų nepakanka, todėl siekiant gauti daugiau žinių iš analizuojamų duomenų, yra naudojami įvairūs duomenų tyrybos (angl. *data mining*) metodai: klasifikavimo, klasterizavimo, vizualizavimo ir kt.

Šio darbo tyrimų sritis yra daugiamačių duomenų skaičiaus mažinimas ir duomenų atvaizdavimas plokštumoje (vizualizavimas).

Viena grupė metodų, įgalinančių atrasti naujas žinias analizuojamose duomenų aibėse, yra vektorių kvantavimo metodai. Vektorių kvantavimas (angl. *vector quantization*) – tai procesas, kurio metu n-mačiai įėjimo vektoriai yra kvantuojami į ribotą aibę n-mačių išėjimo vektorių, kurių skaičius yra mažesnis nei įėjimo vektorių. Dažniausiai šie metodai taikomi garsui ir vaizdui suspausti, tačiau jie tinka ir duomenims klasterizuoti bei klasifikuoti.

Daugiamačių duomenų vizualizavimo, dar kitaip vadinamo dimensijos mažinimo, metodais didelės dimensijos duomenys transformuojami į mažesnės dimensijos erdvę taip, kad išliktų esamos arba būtų atrastos „užslėptos" analizuojamų duomenų savybės. Jais transformavus daugiamačius duomenis į trimatę erdvę ar plokštumą ir juos vizualizavus, daug paprasčiau suvokti duomenų struktūrą ir sąryšius tarp jų. Vienas iš dažnai taikomų metodų yra daugiamačių skalių (angl. *multidimensional scaling*). Daugiamačių skalių metodo vienos iteracijos skaičiavimų sudėtingumas yra $O(nm^2)$, čia m – objektų skaičius, n – dimensijų skaičius, todėl būtina ieškoti būdų skaičiavimams pagreitinti. Vienas iš būdų yra taikyti vektorių kvantavimo metodą ir juo sumažinti duomenų objektų skaičių prieš juos vizualizuojant. Duomenis sumažinti būtina taip, kad

nauja duomenų aibė kaip galima labiau atspindėtų analizuojamos duomenų aibės savybes.

Daugiamačių skalių rezultatas labai priklauso nuo dvimačių vektorių pradinių koordinačių parinkimo būdo, kai daugiamačių skalių paklaida minimizuojama iteraciniu būdu. Siūlomi įvairūs pradinių koordinačių parinkimo būdai, tačiau šio uždavinio sprendimas išlieka aktuali problema.

Šioje disertacijoje sprendžiamos dvi pagrindinės problemos:

1. Duomenų aibės vektorių ir jų dimensijų skaičiaus sumažinimas vektorių kvantavimo ir daugiamačių skalių metodų junginiais, išlaikant duomenų struktūrą;
2. Gautų rezultatų priklausomybės nuo dvimačių vektorių, gautų daugiamačių skalių metodu, pradinių koordinačių parinkimas.

Darbo tikslas ir uždaviniai

Pagrindinis šio darbo tikslas – greitai ir tiksliai atvaizduoti didelės apimties duomenų aibes plokštumoje, tam sukuriant vektorių kvantavimo ir duomenų dimensijų mažinimo metodų junginį ir pasiūlant tinkamus dvimačių vektorių pradinių koordinačių parinkimo būdus.

Siekiant tikslo buvo sprendžiami šie uždaviniai:

- išnagrinėti vektorių kvantavimo strategijas duomenims klasterizuoti;
- ištirti vektorių kvantavimo metodų jungimo galimybes su vizualizavimo metodais, pagrįstais duomenų dimensijų skaičiaus mažinimu;
- ištirti dvimačių vektorių pradinių koordinačių reikšmių parinkimo nuosekliajame junginyje ir integruoto junginio pirmajame mokymo bloke būdus;
- ištirti dvimačių vektorių pradinių koordinačių priskyrimo integruoto junginio visuose mokymo blokuose, išskyrus pirmąjį, būdus;
- sukurti naujus nuoseklaus ir integruoto neuroninių dujų ir daugiamačių skalių metodų junginius, leidžiančius gauti tikslesnę daugiamačių vektorių projekciją plokštumoje ir atlikti išsamią jų lyginamąją analizę su saviorganizuojančio neuroninio tinklo ir daugiamačių skalių junginiais;
- atlikti gautų kvantavimo ir vizualizavimo rezultatų kokybės analizę.

Tyrimo objektas ir metodai

Analizuojant daugiamačius duomenis, norint geriau atskleisti jų struktūrą, vien tik klasikinių vizualizavimo metodų dažnai nepakanka. Disertacijos tyrimo objektai – dirbtiniais neuroniniais tinklais grindžiami vektorių kvantavimo metodai ir daugiamačių duomenų vizualizavimo metodai, pagrįsti dimensijų skaičiaus mažinimu. Su tyrimo objektu betarpiškai yra susiję šie dalykai: daugiamačių duomenų projekcijos į mažesnės dimensijos erdvę kokybės įvertinimo matai, dvimačių vektorių koordinačių parinkimo būdai ir jų atvaizdavimas plokštumoje.

Analizuojant mokslinius ir eksperimentinius pasiekimus daugiamačių duomenų vizualizavimo srityje, buvo naudoti informacijos paieškos, sisteminimo, analizės, lyginamosios analizės ir apibendrinimo metodai.

Remiantis eksperimentinio tyrimo metodu, atlikta statistinė duomenų ir tyrimų rezultatų analizė, kurios rezultatams įvertinti naudotas apibendrinimo metodas.

Darbo mokslinis naujumas

1. Sukurtas nuoseklus neuroninių dujų ir daugiamačių skalių junginys ir integruotas, atsižvelgiantis į neuroninių dujų metodo mokymosi eigą ir leidžiantis gauti tikslesnę daugiamačių vektorių projekciją plokštumoje.
2. Pasiūlyti dvimačių vektorių pradinių koordinačių parinkimo būdai integruoto junginio pirmame mokymo bloke ir koordinačių reikšmių priskyrimo būdai integruoto junginio kituose mokymo blokuose.
3. Eksperimentiškai nustatyta kvantavimo paklaidos priklausomybė nuo neuroninių dujų tinklo mokymo parametrų reikšmių, atliekamų mokymo epochų, neuronų ir neuronų nugalėtojų skaičiaus.
4. Eksperimentiškai ištirta ir parodyta, kad daugiamačių duomenų vizualizavimui neuroninių dujų ir daugiamačių skalių junginys yra tinkamesnis negu saviorganizuojančio neuroninio tinklo ir daugiamačių skalių junginys.

Darbo rezultatų praktinė reikšmė

Tyrimų, atliktų naudojant įvairios prigimties realaus pobūdžio duomenis, rezultatai atskleidė, kad vektorių kvantavimo ir projekcijos mažinimo metodo junginiai gali būti plačiai taikomi daugiamačiams duomenims vizualizuoti. Analizuojant kitus realaus pobūdžio skaitinius duomenis, bus galima remtis išvadomis, gautomis šioje disertacijoje.

Darbo rezultatų aprobavimas

Tyrimų rezultatai publikuoti 8 moksliniuose leidiniuose: 5 periodiniuose recenzuojamuose mokslo žurnaluose, 3 straipsniai konferencijų pranešimų medžiagoje. Tyrimų rezultatai pristatyti šešiose konferencijose Lietuvoje ir užsienyje.

Darbo apimtis

Disertaciją sudaro 5 skyriai ir literatūros sąrašas. Disertacijos skyriai: Įvadas, Vektorių kvantavimo ir vizualizavimo metodai, Vektorių kvantavimo ir projekcijos metodų jungimas, Eksperimentiniai tyrimai, Bendrosios išvados. Disertacijos apimtis 135 puslapiai, juose pateikta 50 paveikslų ir 16 lentelių. Remtasi 81 literatūros šaltiniu.

Bendrosios išvados

Atlikti tyrimai atskleidė dviejų vektorių kvantavimo metodų – saviorganizuojančio neuroninio tinklo ir neuroninių dujų – jungimo su daugiamatėmis skalėmis naujas galimybes. Eksperimentinių tyrimų rezultatai leido daryti šias išvadas:

1. Neuroninių dujų metodu apie 80 % neuronų tampa nugalėtojais, o SOM tik apie 50 %, todėl SOM labiau tinkamas duomenims klasterizuoti. Tačiau kvantavimo paklaida esant tam pačiam neuronų nugalėtojų skaičiui neuroninių dujų metodu visiems analizuotiems duomenims visada mažesnė negu taikant SOM. Neuroninių dujų metodas tinkamesnis daugiamačiams duomenims kvantuoti, o tuo pačiu naudoti junginyje su daugiamačių skalių metodu.
2. Junginyje su daugiamačių skalių metodu naudojant SOM tinklą yra geriau išlaikomi kaimynystės ryšiai tarp taškų, pereinant iš daugiamatės erdvės į dvimatę erdvę Konigo mato prasme negu naudojant neuroninių dujų metodą. Spirmano koeficiento prasme šių abiejų kvantavimo metodų naudojimas yra lygiavertis. MDS paklaida

duomenims, kurių dimensijų skaičius $n = 5, 7, 10$ yra mažesnė, kai junginyje naudojamas SOM, nei ND metodas.

3. Taikant pasiūlytą integruotą neuroninių dujų ir daugiamačių skalių metodų junginį analizuojamų duomenų struktūra jau gerai matoma, kai mokymo blokų skaičius $q < 1/3\,\gamma$, o taikant SOM ir daugiamačių skalių metodo integruotą junginį, mokymo blokų skaičius turi būti $q > 2/3\,\gamma$, čia γ – visų mokymo blokų skaičius.
4. Pasiūlytas pradinių dvimačių vektorių koordinačių priskyrimas pagal vidurinį tašką integruotame junginyje, išskyrus pirmąjį mokymo bloką, gali būti naudojamas kaip alternatyva priskyrimui pagal proporciją, kadangi nepastebėta gautų rezultatų esminių skirtumų.
5. MDS paklaida, gauta nuosekliuoju junginiu, yra mažiausia, kai dvimačių vektorių pradinės reikšmės parenkamos pagal dvi pagrindines komponentes arba dvi didžiausias dispersijas turinčias komponentes, tačiau kartais įmanoma ją dar sumažinti naudojant integruotą junginį. Kai dvimačių vektorių pradinės reikšmės generuojamos atsitiktinai arba parenkamos ant tiesės, tai geriau naudoti integruotą junginį negu nuoseklųjį, nes MDS paklaida mažesnė.

Alma MOLYTĖ

VEKTORIŲ KVANTAVIMO METODŲ JUNGIMO
SU DAUGIAMATĖMIS SKALĖMIS ANALIZĖ

Daktaro disertacija

Fiziniai mokslai (P 000),
Informatika (09 P),
Informatika, sistemų teorija (P 175)

Alma MOLYTE

INVESTIGATION OF COMBINATIONS OF VECTOR QUANTIZATION METHODS WITH MULTIDIMENSIONAL SCALING

Doctoral Dissertation

Physical sciences (P 000),
Informatics (09 P),
Informatics, systems theory (P 175)

2011 05 20 . 1 sp. l. Tiražas 60 egz.
Išleido VšĮ Vilniaus universiteto leidykla
Tauro g. 5, LT-03106 Vilnius.
Interneto svetainė: *http://www.leidykla.vu.lt.*