VYTAUTAS MAGNUS UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Virginijus MARCINKEVIČIUS

INVESTIGATION AND FUNCTIONALITY IMPROVEMENT OF NONLINEAR MULTIDIMENSIONAL DATA PROJECTION METHODS

Summary of Doctoral Dissertation

Physical Sciences, Informatics (09 P)

Vilnius, 2010

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2003–2010.
Scientific Supervisor:

Prof Dr Habil Gintautas DZEMYDA (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

This dissertation is being defended at the Council of Scientific Field of Informatics at Vytautas Magnus University:

Chairman:

Prof Dr Habil Vytautas KAMINSKAS (Vytautas Magnus University, Physical Sciences, Informatics – 09P).

Members:

Prof Dr Habil Juozas AUGUTIS (Vytautas Magnus University, Physical Sciences, Informatics – 09P),

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

Prof Dr Habil Romualdas BAUŠYS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Dr Julius ŽILINSKAS (Institute of Mathematics and Informatics, Physical Sciences, Informatics – 09P).

Opponents:

Prof Dr Habil Mifodijus SAPAGOVAS (Institute of Mathematics and Informatics, Physical Sciences, Informatics – 09P),

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

The dissertation will be defended at the public meeting of the Council of Scientific Field of Informatics in the auditorium number 203 of the Institute of Mathematics and Informatics at 1 p. m. on 28 September 2010.
Address: Akademijos str. 4, LT-08663 Vilnius, Lithuania.
Tel.: +370 5 210 9300, fax +370 5 272 9209;
e-mail: mathematica@ktl.mii.lt
The summary of the doctoral dissertation was distributed on 24th of August 2010.
A copy of the doctoral dissertation is available for review at the Library of Vytautas Magnus University (K. Donelaičio str. 58, LT-44248 Kaunas, Lithuania) and at the Library of Institute of Mathematics and Informatics (Akademijos str. 4, LT-08663 Vilnius, Lithuania).

VYTAUTO DIDŽIOJO UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Virginijus MARCINKEVIČIUS

NETIESINĖS DAUGIAMAČIŲ DUOMENŲ PROJEKCIJOS METODŲ SAVYBIŲ TYRIMAS IR FUNKCIONALUMO GERINIMAS

Disertacijos santrauka

Fiziniai mokslai, informatika (09 P)

Vilnius, 2010

Disertacija rengta 2003–2010 metais Matematikos ir informatikos institute.
Darbo mokslinis konsultantas:

prof. habil. dr. Gintautas DZEMYDA (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija ginama Vytauto Didžiojo universiteto Informatikos mokslo krypties taryboje:

Pirmininkas:

prof. habil. dr. Vytautas KAMINSKAS (Vytauto Didžiojo universitetas, fiziniai mokslai, informatika – 09P).

Nariai:

prof. dr. habil. Juozas AUGUTIS (Vytauto Didžiojo universitetas, fiziniai mokslai, informatika – 09P),

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

prof. habil. dr. Romualdas BAUŠYS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

dr. Julius ŽILINSKAS (Matematikos ir informatikos institutas, fiziniai mokslai, informatika – 09P).

Oponentai:

prof. habil. dr. Mifodijus SAPAGOVAS (Matematikos ir informatikos institutas, fiziniai mokslai, informatika – 09P),

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

Disertacija bus ginama viešame Informatikos mokslo krypties tarybos posėdyje 2010 m. rugsėjo mėn. 28 d. 13 val. Matematikos ir informatikos instituto 203 auditorijoje.
Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.
Tel.: +370 5 210 9300, fax +370 5 272 9209;
el.paštas: mathematica@ktl.mii.lt
Disertacijos santrauka išsiuntinėta 2010 m. rugpjūčio 24 d.
Disertaciją galima pažiūrėti Vytauto Didžiojo universiteto (K. Donelaičio g. 58, LT-44248 Kaunas, Lietuva) ir Matematikos ir informatikos instituto (Akademijos g. 4, LT-08663 Vilnius, Lietuva) bibliotekose.

Introduction

Relevance of The Problem

Data comprehension is a difficult process, especially if that data refer to complicated object or phenomenon, that is characterized by various quantitative and qualitative parameters or features. That kind of data is called multidimensional data and may be interpreted as points or position vectors in multidimensional space. To analyze multidimensional data we often use one of the main instruments of data analysis – data visualization or graphical presentation of information. The fundamental idea of visualization is to provide data in the form that would let the user to understand the data, to make conclusions, and to influence directly the further process of decision making. Visualization allows better comprehension of complicated data sets, may help to determine their subsets that interest the researcher. Dimension reducing methods allow discarding interdependent data parameters, and by means of projection methods it is possible to transform multidimensional data to a line, 3D space or other form that may be comprehended by a human eye. It is much quicker and easier to comprehend visual information than numeric or textual. On the other hand, that kind of comprehension may suit only as a ground for hypothesis and further research held using strict mathematical models. What information and how it should be visualized depends on the user working in this field, thus there arise some problems that must be solved: what visualization methods to choose and how to select the optimal parameters. As a result of constantly increasing data sets we have more and more multidimensional data visualization methods, however the relevant problem remains – approval of those methods, and researches of validity of their application.

The Object of Research

The object of the research done in this dissertation is multidimensional data, the presentation of that data by nonlinear multidimensional scaling and self-organizing maps, and evaluation of projection quality.

The Objective and Tasks of the Thesis

The objective of this work is to improve the functionality of nonlinear multidimensional data projection methods by examining their characteristics.

Aiming for that objection the following tasks were accomplished:

1. To examine data initialization methods for multidimensional scaling algorithm.
2. To compare multidimensional scaling SMACOF algorithm, Sammon's mapping algorithm, and relative multidimensional scaling algorithm by criterions that evaluate topology preserving.
3. To examine the effectiveness of diagonal majorization algorithm by comparing it with multidimensional scaling SMACOF realization, and with relative multidimensional scaling algorithm.
4. To examine theoretically the numerical dependence of neurons-winners on the training epoch in the self-organizing map (SOM).
5. To research new possibilities to represent the SOM.
6. To modify relational perspective map algorithm with the view to improve its convergence.

Scientific Novelty

Research made in this work revealed new possibilities to develop multidimensional data visualization methods and instruments.

There was proved that the initial selection of projection data on the line in Sammon's mapping algorithm is inexpedient. Regarding this, it is advisable to use principal component analysis (PCA) or the largest dispersion method to select initial points.

There was shown that the effectiveness of diagonal majorization algorithm resigns to the multidimensional scaling SMACOF realization, and the relative multidimensional scaling.

There was theoretically examined the dependence of number of the neurons, recalculated in one epoch in the SOM of rectangular form, on the training epoch.

The work introduced a new way to represent neural SOM net: the color of cells in neural net table is selected as a tone of grey that relates to the length of codebook vector corresponding to neuron in the cell.

There was offered a new way to select initial data for the multidimensional scaling algorithms upon the largest dispersions.

Relational perspective map algorithm was examined, and the author offered to use two new functions of distance, reassuring RPM algorithm convergence.

Methodology of the Research

To analyze scientific and experimental achievements in the field of data visualization there were used search of information, systematization, analysis, comparative analysis, and generalization methods.

Software engineering method was used to create software. To prove theorems and to examine convergence of algorithms, theoretical examination methods were used. To prove statements, mathematic induction principal was applied.

Referring to experimental examination method, the statistical analysis of data and examination results was made. To evaluate its results, there was generalization method applied.

Practical Significance of Achieved Results

Research results were applied on projects supported by the Lithuanian State Science and Studies Foundation and the Research Council of Lithuania:

- "Information technologies for human health – support for clinical decisions (eHealth), IT health (No. C-03013)". Start date: 09-2003; finish date: 10-2006.
- High technologies development program project "Atherosclerosis pathogenesis peculiarities determined by human genome variety peculiarities (AHTHEROGEN) (No. U-04002)". Start date: 04-2004; finish date: 12-2006.
- "Information technology tools of clinical decision support and citizens wellness for e.Health (No. B-07019)". Start date: 09-2007; finish date: 12-2009.
- Underlying Lithuanian scientific research and experimental development direction project "Genetic and genomic lip and (or) palate non-union basis research (GENOLOG) (No. C-07022)". Start date: 04-2007; finish date: 12-2009.
- Integrated work program of Lithuania and France in the field of bilateral cooperation scientific research and experimental development "Žiliberas" (No. V-09059). Start date: 04-2008; finish date: 12-2010.

The Defended Statements

1. Initiation of projection data on the line in Sammon's mapping algorithm is inexpedient, because the convergence of error in iteration process is slow.

2. Diagonal majorization algorithm (DMA), in relation to error, yields to the multidimensional scaling SMACOF realization and the relative multidimensional scaling algorithm. DMA is faster than SMACOF, however DMA error is bigger than that of SMACOF or relative multidimensional scaling algorithm.
3. In the SOM of rectangular form with the largest edge of k' neurons, the number of retrained neurons is of a staircase form and decreases while the training epoch order number is increasing – it decreases by one after $e' = \left[\frac{n'e}{k'}\right] - \left[\frac{(n'-1)e}{k'}\right]$ $(n' = 1, \dots, k' - 2)$ epochs.
4. It is possible to apply new functions of distance, which greatly improves the performance of the relational perspective map algorithm.
5. The choosing of initial points according to the biggest dispersions in the multidimensional scaling algorithms is one of the most precise and effective ways to select them.

The Scope of the Scientific Work

The volume of work is 105 pages; are used 57 numbered formulas, 29 figures, and 13 tables in the text. The thesis lists 107 references. The dissertation consists of introduction, 3 chapters, conclusions, and the list of references.

1. Nonlinear Multidimensional Data Projection Methods

The chapter is devoted to review multidimensional scaling methods, such as Sammon's mapping, SOM, DMA, relative multidimensional scaling and RPM. All methods have been investigated in the thesis. Some theoretical results, obtained in the investigation, are presented in the chapter. Multidimensional scaling (MDS) is a group of methods that project multidimensional data to a low (usually two) dimensional space and preserve the interpoint distances among data as much as possible. Let us have vectors $X_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$, $i = \overline{1,s}$ $(X_i \in R^n)$. The pending problem is to get the projection of these *n*-dimensional vectors X_i, $i = \overline{1,s}$ onto the plane R^2. Two-dimensional vectors $Y_1, Y_2, \dots, Y_s \in R^2$ correspond to them. Here $Y_i = \{y_{i1}, y_{i2}\}$, $i = \overline{1,s}$. Denote the distance between the vectors X_i and X_j by δ_{ij}, and the distance between the corresponding vectors on the projected space (Y_i and Y_j) by d_{ij}. In our case, the initial dimensionality is n, and the resulting one is 2. There exists a multitude of variants of MDS with slightly different so-called stress functions. In our experiments, the raw stress is minimized:

$E_w(Y) = \sum_{i,j=1,i<j}^{s} w_{ij}\left(\delta_{ij}(X) - d_{ij}(Y)\right)^2$, where w_{ij} are weights. The Guttman majorization algorithm based on iterative majorization (SMACOF) is one of the best minimization of the stress function algorithms for this type of minimization problem. This method is simple and powerful, because it guarantees a monotone convergence of the stress function.

Formula $Y(m'+1) = V^{+}B\big(Y(m')\big)Y(m')$, is called the Guttman transform. Where matrix $B\big(Y(m')\big)$ and matrix of weights V have the entries:

$$b_{ij} = \begin{cases} -\frac{w_{ij}\delta_{ij}}{d_{ij}}, & when\ i \neq j\ and\ d_{ij} \neq 0; \\ 0, & when\ i \neq j\ and\ d_{ij} = 0; \\ -\sum_{j=1,j\neq i}^{n} b_{ij}, & when\ i = j. \end{cases} \text{ and } V = \begin{pmatrix} \sum w_{1j} & & & \\ & \ddots & -w_{ij} & \\ & -w_{ij} & \ddots & \\ & & & \sum w_{sj} \end{pmatrix}$$

V^{+} denotes the Moore-Penrose pseudoinverse of V. m' is an iteration number.

Diagonal majorization algorithm uses simpler majorization function: $Y(m'+1) = Y(m') + \frac{1}{2}diag(V)^{-1}[B(Y(m') - V)]Y(m')$.

DMA attains slightly worse projection error than SMACOF, but computing by the iteration equation is faster and there is no need of computing the pseudoinverse V^{+} matrix. In addition, DMA differs from SMACOF that a large number of entries b_{ij} in matrix V have zero values. This means that iterative computations of two-dimensional coordinates, $Y_i, i = \overline{1,s}$ are based not on all distances δ_{ij} between multidimensional points X_i and X_j. This allows to speed up the visualization process and to save the computer memory essentially.

This algorithm, however, remains of $O(s^2)$ complexity if we use the standard V matrix. With a view to diminish the complexity of this algorithm, is used only a part weights of matrix V. The weights are defined by setting $w_{ij} = 1$ for k "cycles" of δ_{ij}, e.g., $i \leftrightarrow i \pm 1, \dots, i \leftrightarrow \pm k$, etc. and $w_{ij} = 0$, otherwise. The parameter k defines neighbourhood order for point X_i in the list of analysed data set vectors.

Knowing the dimension s and parameter k of quadratic matrix, we find that complexity of DMA algorithm is $O(s^2 - (s-2k-1)^2)$, when $s \geq 2k+1$. When all weights are used, the complexity of DMA algorithm is $O(s^2)$. Hence DMA calculation time, dependently on the k selected, shortens proportionally as, $s^2/(s^2 - (s-2k-1)^2)$, here $s \geq 2k+1$. When s are large enough and $k = s/10$ calculation time shortens to 2.77 times. When $k = s/100$, calculation time shortens to 25.25 times.

Willing to determine how much the projection result depends on the parameter k, and what is optimal value for this parameter, the research was made. This research is presented in the Experimental Research chapter.

This chapter examines RPM algorithm. It visualizes multidimensional data onto the closed plane (torus surface) so that the distances between data in the lower-dimensional space would be as close as possible to the original distances. But what is more important, the RPM method also gives the ability to visualize data in a non-overlapping manner so that it reveals small distances better than other known visualization methods.

From the physical point of view, the torus is a force directed multiparticle system: the image points are considered as particles that can move freely on the surface of the torus, but cannot escape the surface. The particles exert repulsive forces on one another so that, guided by the forces, the particles rearrange themselves to a configuration that visualizes the relational distances δ_{ij} . While mapping data points on a torus, the RPM algorithm minimizes the potential energy: $E_p = \sum_{i<j}(\delta_{ij}/pd_{ij}^p)$, when $p = -1, p > 0$. When $p = 0$ than $E_0 = -\sum_{i<j}\delta_{ij}\, ln\big(d_{ij}\big)$. Here the parameter p is called the rigidity.

Distances on torus $T = [0, w] \times [0, h] \subset R^2$ are calculated using formula: $d_r(Y_i, Y_j) = \big(min\{|y_{i1} - y_{j1}|, w - |y_{i1} - y_{j1}|\}^r + min\{|y_{i2} - y_{j2}|, h - |y_{i2} - y_{j2}|\}^r\big)^{\frac{1}{r}}$, here w and h is width and height of surface of torus, parameter r is usually 2.

E_p is minimized applying iterative Newton-Rapson method, but using function d_r of distances, function E_p of error is not differentiated, when $E_p|y_{i1} - y_{j1}| = \frac{w}{2}$ or $y_{i1} = y_{j1}$ (the same with another coordinates).

RPM algorithm. The other problem is that RPM algorithm not converges at all, when selected values of w and h are very different. It may be explained by the fact that, even if coordinates y_{i1} and y_{i2} of point Y_i are calculated individually, they influence the value of one another. Calculating the distance d_{ij} both coordinates are evaluated. If influence of one of them is much stronger (this happens when $w \gg h$ or $w \ll h$), then disproportionate influence of vector coordinates y_{i1} and y_{i2} on the value of one another is inevitable. This problem in the work is partially solved using distance d_n, which is obtained by normalizing distance d_0 and is equal to $d_n\big(Y_i, Y_j\big) = min\left\{\frac{|y_{i1}-y_{j1}|}{w}, 1 - \frac{|y_{i1}-y_{j1}|}{w}\right\} + min\left\{\frac{|y_{i2}-y_{j2}|}{h}, 1 - \frac{|y_{i2}-y_{j2}|}{h}\right\}$. The problem is solved only partially, because applying normalized distance d_n, RPM algorithm doesn't depend on torus parameters w and h, and without additional stopping parameters we get projection, however this projection isn't stable. Points, situated near the surface

of torus, leap from one side of the torus to another, thus moving all other projection points. Applying continuous function for distance: $d_n(Y_i, Y_j) = min\left\{\frac{|y_{i1}-y_{j1}|}{w}, 1-\frac{|y_{i1}-y_{j1}|}{w}\right\} + min\left\{\frac{|y_{i2}-y_{j2}|}{h}, 1-\frac{|y_{i2}-y_{j2}|}{h}\right\}$, E_p becomes differentiable on all points of torus surface. However it doesn't change convergence of function E_p.

Function of distance d_t was deduced with reference to partial derivatives of distance d_0. Since partial derivates can be approximated by sinusoid, the new distance d_t may be derived with reference to full differential of function. Using this function of distance, in points where E_p wasn't differentiated with other functions of distance, it is equal to zero. Thus locations of points near the surface of torus vary marginally and gradually.

If we want stable convergence of E_p to the point of minimum, we need to consider E_p as function of two variables $f(y_{i1}, y_{i2})$, not one.

The self-organizing map. The SOM is a class of neural networks that are trained in an unsupervised manner, using competitive learning. It is a well-known method for mapping a high-dimensional space onto a low-dimensional one. We consider here a mapping onto a two-dimensional grid of neurons. Usually, the neurons are connected to each other via a rectangular or hexagonal topology. The rectangular SOM is a two-dimensional array of neurons $M = \{m_{ij},\ i = 1, \dots, k_x, j = 1, \dots, k_y\}$.

Here k_x is the number of rows, and k_y is the number of columns. Each component of the input vector is connected to every individual neuron. Any neuron is entirely defined by its location on the grid (the number of row i and column j) and by the codebook vector, i.e., we can consider a neuron as an n-dimensional vector $m_{ij} = \{m_{ij}^1, m_{ij}^2, \dots, m_{ij}^n\} \in R^n$.

The learning starts from the vectors m_{ij} initialized randomly. At each learning step, an input vector X_i is passed to the neural network. The Euclidean distance from this input vector to each vector m_{ij} is calculated and the vector (neuron) m_{ij}^c with the minimal Euclidean distance to X_c is designated as a winner. The components of the vector m_{ij} are adapted according to the rule: $m_{ij} \leftarrow m_{ij} + h_{ij}^c(X_i - m_{ij})$, where $h_{ij}^c = \frac{\alpha}{\alpha\eta_{ij}^c+1}$, $\alpha = \max\left(\frac{e+1-\hat{e}}{e}, 0{,}01\right)$; h_{ij}^c is the neighbourhood order between the neurons m_{ij} and m_{ij}^c (all neurons adjacent to the given neuron can be defined as its neighbours of a first order, then the neurons adjacent to a first-order neighbour, excluding those already considered, as neighbours of a second order, etc.); e is the number of training epochs; $\hat{e}$ is the order number of a current epoch $\hat{e} \in [1, e]$. We recalculate the

vector m_{ij} if $\eta_{ij}^c \le max\left[\alpha\, max\left(k_x, k_y\right), 1\right]$. Let us introduce a term "training epoch". An epoch consists of *s* steps: the input (analysed) vectors from X_1 to X_s are passed to the neural network in a consecutive or random order.

A theorem about the SOM training has been formulated and proved. Denote $k' = max\left(k_x, k_y\right)$. It follows from the rule of SOM training that $\eta_{ij}^c = k'$, as $\hat{e} = 1$. n' is the integer number that indicates how much the neighbourhood order has been decreased as compared with the maximal one (k'). Then the following theorem is valid.

Theorem 1. If we have rectangular SOM net, the edge of which is $k' = \max\left(k_x, k_y\right) \le 100$, and training epoch answer inequality $1 \le \hat{e} \le e + 1 - e/k'$, after the epoch, whose number is $\hat{e} = \left[\frac{(n'-1)e}{k'}\right] + 2$, $(n' = 1, \dots, k' - 1)$, the maximal neighbourhood order η_{ij}^c of any neuron m_{ij}^c is lower than that after the $(\hat{e} - 1)$-st epoch by one, if $1 \le \hat{e} \le e + 1 - e/k'$. The maximal neighbourhood order does not decrease and remain equal to one $(\eta_{ij}^c = 1)$ for $e + 1 - e/k' \le \hat{e} \le e$.

If follows from this theorem that dependence of the number of retrained neurons in SOM net on the order number is of a staircase form and decreases after each $e' = \left[\frac{n'e}{k'}\right] - \left[\frac{(n'-1)e}{k'}\right]$ $(n' = 1, \dots, k' - 2)$, numbers of epochs.

In the case of the rectangular topology (k_x rows and k_y columns), we can draw a table with cells corresponding to the neurons. However, the table and its properties do not answer the question, how much the vectors of the neighbouring cells are close in the n-dimensional space. The answer may be found, by using additional visualization of SOM, for example, graphic display, called the U-matrix (Unified distance matrix), component planes, a histogram. The U-matrix that illustrates the clustering of codebook vectors in the SOM has been developed by Ultsch, Siemon and Kraaijveld. They have proposed a method in which average distances between the neighbouring codebook vectors are represented by shades in a grey scale (or, eventually, pseudo-color scales might be used). If the average distance of neighbouring neurons is short, a light shade is used; dark shades represent long distances: high values of the U-matrix indicate a cluster border; uniform areas of low values indicate the clusters themselves (Fig. 1, left-hand side). In Fig. 1 left-hand side, the U-matrix is presented. Iris data are analysed. It is known that the first iris kind (1 – Iris Setosa) forms a separate cluster; the second kind (2 – Iris Versicolor) and the third one (3 – Iris Virginica) are mixed a little bit (Fig. 1).

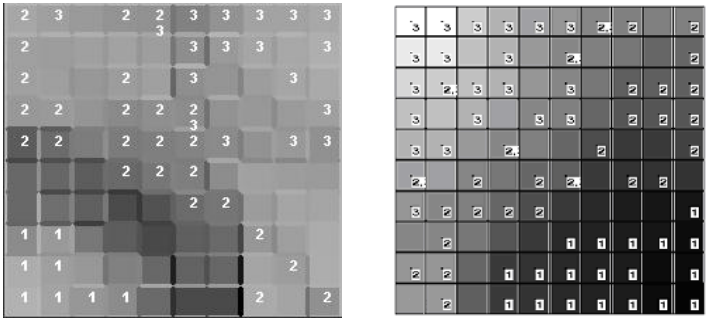

Fig. 1. Examples of the SOM visualization: U-matrix and neuron-vector length-based visualization

In this section, a new way of the result visualization is suggested – a *neuron-vector length-based visualization* (Fig. 1, right-hand side). Directions of neuron-vectors in the neighbouring cells are similar in the trained SOM. We notice that the lengths of neuron-vectors have some specific distribution: the similarity of the neuron-vectors may be estimated not only in accordance with their directions, but also with their lengths. The cells of the SOM are painted, using the different grey shading. Intensity of the cell color is proportional to the length of vectors. A darker color means shorter vectors. Another way is to put the number of the analysed input vectors, related with the corresponding vectors-winners into the cells that correspond to the vectors-winners. It would allow drawing conclusions on the nearness of the analysed vectors X_i, $i = \overline{1,s}$, their clusters and densities of the distribution of the vectors. When comparing both sides of Fig. 1, both of them allow us to draw similar conclusions on the clusters of the analysed vectors. However, the results by the neuron-vector length-based visualization seem clearer.

2. Research Methodology

As computer software is rapidly developing, algorithm computation time is shortening; also the schemes of the algorithms themselves are changing. Thus it is possible to examine algorithms in larger and larger data sets, also such algorithms, that require to do more operations in computer processors. This led to the selection of different size data sets: from “Fisher Iris” (150x4) to “Satimage” (6435x36).

This chapter analyzes the problems of initialization of data, used in the research, before rendering it to visualization algorithms. It also gives theoretical results, solving Sammon algorithm initial vectors initialization problem.

The initial values of projection vectors $Y_i \epsilon R^d$ influence a final result in nonlinear projection methods. Optimization methods, used in visualization algorithms, often find a local, but not the global, optimum of a function that characterizes the quality of projection. For this reason, location of the initial vectors is very important, i.e., different local optima are often obtained for different sets of initial vectors.

The projection vectors $Y_i \in R^d$, $d = 1,2,3$ may be initialized in various ways. One of the simplest ways is a generation of initial vectors at random in some area. The shape of that area usually is a square or a cube, but some other forms like a line, a plane, a sphere, *etc.*, are also possible. In this case, a projection algorithm is repeated a lot of times with different sets of initial vectors Y_i, and the most faithful mapping, corresponding to the best found value of the mapping criterion (e.g. minimal projection error), is selected as the final one. In SOM_PAK software, this way is also applied with a slight modification: the first coordinate lies on a line, and the second one is selected at random. This is an empirical, theoretically ungrounded result.

However, such an initialization way is unreasonable, it requires more computing time. Therefore other initialization ways could be used.

Let's analyze theoretically, how, in the case of d = 2, the projection of points is changing, if initial vectors $Y_i = \{y_{i1}, y_{i2}\}$ are initiated on the line $y_{i1} = a \cdot y_{i2} + b$, here a and b are some real number constants. For initiation on the line Sammon's mapping was examined.

Sammon's mapping is a nonlinear projection method, which is closely related to the metric MDS version described above. It also tries to optimize the cost function (*Sammon stress*) that describes how well the pair wise distances in a data set are preserved. The cost function of Sammon's mapping is the following distortion of projection: $E_S = \frac{1}{\sum_{i,j=1,i<j}^{s} \delta_{ij}} \sum_{i,j=1,i<j}^{s} \frac{(\delta_{ij} - d_{ij})^2}{\delta_{ij}}$. In Sammon's mapping, the dissimilarity of vectors X_i, X_j and Y_i, Y_j to d_{ij} and δ_{ij} accordingly are evaluated as distances between the coordinates both in projection and input space R^n. These distances can be calculated using any metrics, but Sammon suggest using Euclidian metric. Sammon stress function value is more sensitive to small distances than to large ones.

Iterative gradient pseudo–Newton method, based on diagonal approximation of Hessian matrix, is used to minimise error in Sammon's mapping. The coordinates y_{ik}, $i = \overline{1,s}$, $k = 1,2$ of projection vectors $Y_i \in R^2$

are computed by iteration formula: $y_{ik}(m'+1) = y_{ik}(m') - \alpha \cdot \frac{\partial E_S(m')}{\partial y_{ik}(m')} / \left|\frac{\partial^2 E_S(m')}{\partial y_{ik}^2(m')}\right|$, here m' denotes the iteration order number; α is a parameter, which influences the optimization step. J. W. Sammon called it a "magic factor".

*Theorem 2***.** If the initial points $Y_i = \{y_{i1}, y_{i2}\}$, $i = \overline{1,s}$ for Sammon's mapping are located on the line $y_{i1} = a \cdot y_{i2} + b$ ($a = \pm 1, b \in R$), then the projection of points, calculated by Sammon stress, will be located on the same line.

*Conclusion***.** It follows from the theorem 2 that the points will always stay on the line. However, it has been proved experimentally that the coordinates of two-dimensional points marginally vary on the line in the first iterations; but after several iterations, the points deviate from the line, disperse onto the plane. The reason for that is inevitable computation errors.

Therefore, this way of initialization (on a line) is possible, but the initial iteration process is very slow. It is necessary to look for the ways of accelerating this process.

A more complicated way of initialization is the use of PCA. At first, multidimensional data are projected on the plane using PCA, two-dimensional points are obtained; then, namely these points are set as the initial two-dimensional points. However, search for the principal components is a complicated time-consuming problem.

We suggest a simpler way: to calculate the variances of each j-th component of $X_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$, using s values and select the coordinates of the initial two-dimensional vectors to be equal to the values of two parameters the dispersion of which are largest. Let us call it a by dispersion method.

Quantative criteria of mapping. The problem of objective comparison of the mapping results arises when the multidimensional data are visualized using various methods that optimize different criteria of the mapping quality. It is necessary to select a set of universal criteria that describe the projection quality and may be general for different methods. Minimal wiring (MW) coefficient and Spearmen coefficient (rho) were used for this purpose in the dissertation.

Computer hardware, used for the research, and specification of software created are introduced in this chapter. The need to create such software emerged when it was necessary to consolidate various multidimensional scaling algorithms realizations to one system. Furthermore, the software must have: defined structure, hierarchies of classes and data types, that software available may be easy supplemented by new functions and visualization algorithms. The other problem is to make sure that the software will work in various operating systems like *Unix, Linux or Windows*. It is required that software code would

be as universal as possible in regard to its compilation by various compilers. To meet those requirements it was chosen to apply C programming language, and it is advised to move graphical user interface to internet server, created using HTML, JavaScript, PHP technologies. Working example of such software, intended to make experiments in computers cluster, is accessible through *http://cluster.mii.lt/visualization*.

3. Experimental Research

This chapter gives substantiation by experiment of means to select parameters of particular multidimensional data visualization methods, analyzed in the dissertation.

Investigation of the Sammon and SMACOF algorithms. This chapter analyzes optimization of multidimensional data representation. Sammon projection, multidimensional scaling SMACOF realization and consequent combinations of them with the SOM net, using distances, computed by Euclidian metrics, are examined. In this research, the characteristics of algorithms and their combinations are analyzed. The methods to select initial vectors are examined; their comparative analysis is made; they are assessed by different quantitative criterions. Quantitative criterions allow to evaluate the results of projection, and to choose the best. The research was made using data sets of six different origins. It has been proved experimentally that the examined realization by the SMACOF uses approximately 2.3 times less computing time than the realization of Sammon's mapping for the same number of iterations, for a sufficiently large number of iterations. Here, one iteration contains calculations, where both components of all the two-dimensional points are recalculated. The reason is that Sammon's mapping requires more complex calculations as compared with the SMACOF.

The combinations SOM_Sammon or SOM_MDS are examined using various data sets. The results of the SOM training quality depend on the initial values of the neurons-vectors $m_{ij} = \{m_{ij}^1, m_{ij}^2, \dots, m_{ij}^n\}$. Therefore, it is advisable to train the SOM several times, using different sets of the initial neurons-vectors, and to choose such a trained map that the SOM error $E_{SOM} = 1/s \sum_{c=1}^{s} \|X_c - m_{ij}^c\|$ were the least. The experiments have been repeated for 100 times and a set of vectors-winners that corresponds to the least SOM error E_{SOM}, was chosen. Then the vectors-winners were visualized using Sammon's or MDS algorithms. In the experiments, the number of iterations of Sammon's and MDS algorithms has been chosen so that the computing time of both methods be approximately equal.

Their projections have been obtained using SOM_Sammon or SOM_MDS combinations. The values of mapping quality criteria have been calculated (Table 1).

Table 1. Values of various criteria, obtained using SOM_Sammon or SOM_MDS combinations

Criteria	Type	SOM_Sammon					
		"Iris"	"HBK"	"Wood"	"Wine"	"Cancer"	"Cluster"
MW	decrease	**21.94382**	**4.23492**	**1.35353**	**88.19783**	164.0650	**44.15650**
Spearmen coefficient	increase	**0.99664**	**0.98705**	**0.95675**	**0.98805**	**0.98310**	0.83153
Criteria		SOM_SMACOF					
MW	decrease	20.92690	3.95750	1.27246	86.29917	**181.1525**	35.78920
Spearmen coefficient	increase	0.99864	0.99026	0.96069	0.98919	0.98318	**0.81109**

In Table 1, the "type" shows how the measure changes with an increase in the mapping quality, so that for the type "decrease" small numbers mean better maps. Here numbers in bold indicates a better result.

Table 1 shows that the quality of maps, obtained by SOM_MDS algorithm, is better as compared with the maps, obtained by SOM_Sammon, in many cases. However, the difference between the values of criteria is insignificant, therefore projection mappings are similar. Therefore both combinations can be used in the visualization of multidimensional data with a sufficiently good quality.

Investigation of DMA algorithm. In DMA algorithm, parameter k defines neighbourhood order for point X_i in the list of analysed data set vectors.

Selection of the parameter k in the DMA algorithm has a great influence on the projection error $E = \sqrt{\sum_{i,j=1,i<j}^{s} w_{ij}\left(\delta_{ij} - d_{ij}(Y)\right)^2 / \sum_{i,j=1,i<j}^{s} w_{ij}\delta_{ij}^2}$ and obtained map. It has been investigated how the projection error is varying by increasing the parameter k, the computing time and number of iterations being fixed. The vectors of the initial analysed data set were mixed at random in each experiment so that there were less similar points in the list of analysed data points. Having done 50 experiments for each k, when k varied from 100 to 1000, by step 100, the averages of errors were computed. The initial two-dimensional vectors were initiated in SMACOF and DMA algorithms by the method of PCA.

The experiment has shown (see Fig. 2 and Fig. 3) that, for and under the fixed computing time, already after 300 iterations one can get quite an accurate result. Projection error increases less than 1 % comparing with SMACOF algorithm. With an increase in the number of iterations, the error changes but slightly. By increasing k considerably, the computing time also increases, while the result approaches that obtained by SMACOF algorithm ($E = 0.043497$ for "Abalone" data set and $E = 0.210109$ for "Ellipsoidal" data set).

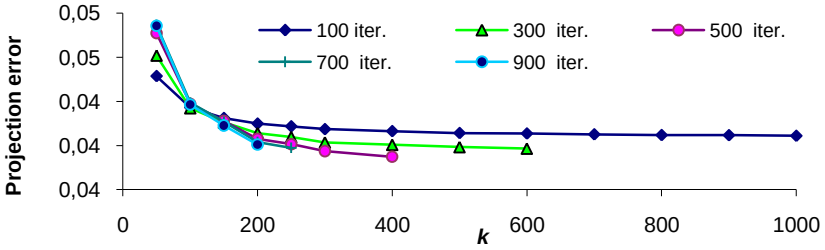

Fig. 2. Dependence of the projection error on the neighbourhood order parameter ***k*** (for "Abalone" data set)

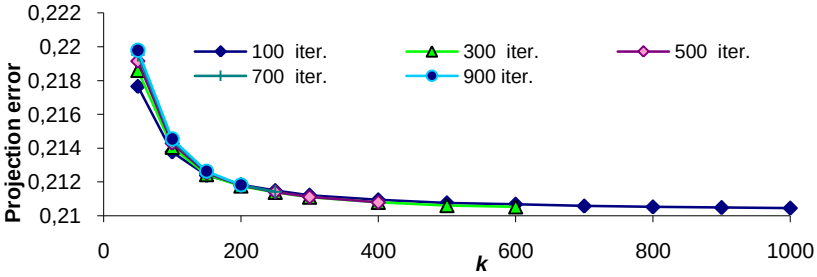

Fig. 3. Dependence of the projection error on the neighbourhood order parameter ***k*** (for "Ellipsoidal" data set)

This experiment has also illustrated that for too small k increasing number of iterations, in many cases the error does not decrease but, vice verse, increases (see Fig. 2 and Fig. 3, $k = 50$).

Carrying out the experiments with different data sets, it has been established that the projection error is influenced a great deal by formation of set of multidimensional points, i.e., numbering of vectors in analysed data set. To corroborate this fact, the following investigation was performed. The initial set

of multi-dimensional data was made up using three different strategies of points numbering of the set:

Strategy I. At the beginning of algorithm operation, the points of analysed multidimensional data set are mixed up at random (one random numbering).

Strategy II. The points of multidimensional data set and two-dimensional vectors, corresponding to these multidimensional points, and whose coordinates have been calculated in the previous iterations, are randomly mixed up in the operation of the algorithm at the beginning of the each iteration (random numbering before each iteration).

Strategy III. Using the method of the PCA, multidimensional vectors are projected onto a straight line, thus establishing the similarity of this point, and multi-dimensional data are numbered in this order (closer points should have similar order numbers).

Using Strategies I and II for multidimensional vector numbering, 50 experiments have been done with each k, varying it from 50 to 1000, the data have been visualized, the averages of projection error and standard deviation as well as computing time has been recorded. Since the previous experiments have shown that the error changes insignificantly after more than 300 iterations, the algorithms have been iterated 300 times each in this experiment (Strategy I is used, Fig. 2 and Fig. 3).

Using Strategy II, even after 100 iterations rather good results have been obtained and by increasing the number of iterations they almost do not change. Using this strategy, the least error is obtained, when these three strategies were compared. The projection error varies insignificantly by increasing k (Fig. 4, Fig. 5 and Fig. 6). Increasing parameter k from 300 to 1000, the projection error decreases by the rule $E = -0.0002 \ln(k) + C$, here C is a constant. It means that in this case, the projection error will be decreased till 0.08 % approximately.

The experiments done have illustrated that numbering of multidimensional data (Strategy III) worsens the visualizations results (Fig. 4, Fig. 5). If we employ the DMA algorithm, we need close and distant points side by side, because taking them into consideration the coordinates of two-dimensional vectors are computed. Mixing of multidimensional vectors at each iteration implies that when calculating the coordinates of a two-dimensional point, more and various neighbours are regarded, which results in a more accurate projection (Strategy II) and it suffices less iterations (100 is enough) (Fig. 5).

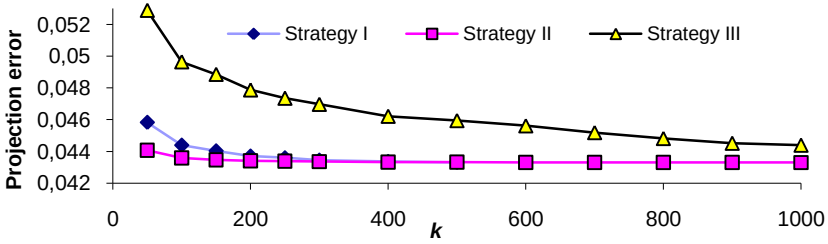

Fig. 4. Dependence of the projection error on the neighbourhood order parameter ***k*** (for "Abalone" data set), using different numbering strategies

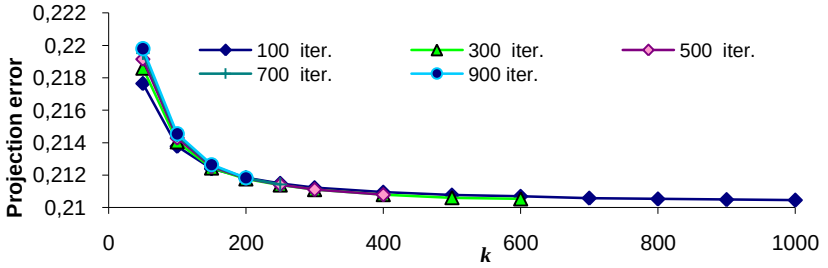

Fig. 5. Dependence of the projection error on the parameter ***k*** (for "Ellipsoidal" data set), using different numbering strategies

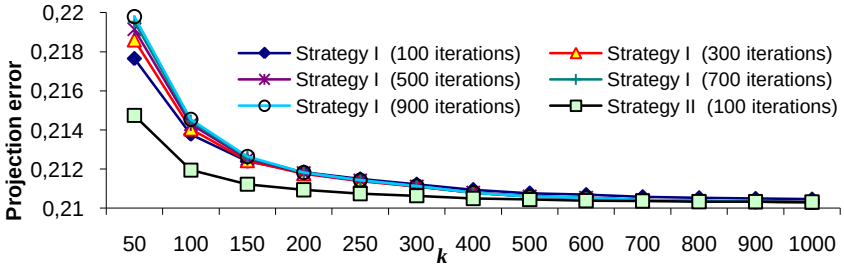

Fig. 6. Dependence of the projection error on the parameter ***k*** ("Ellipsoidal" data set), using different numbering strategies (Strategies I and II) and different number of iterations

Also the SMACOF and DMA algorithms have been compared with respect to time and projection error. After the experiments with four different

data sets, it has been established that the projection error, obtained by SMACOF, is slightly smaller, while using DMA, the computing time is considerably shorter. The larger the set, the more distinct the computing time difference is. By comparing visualization results obtained by SMACOF and DMA, we notice no great difference between the obtained projections, since the difference between errors is very small (≤ 1 % for "Abalone", "Gaussian" and "Ellipsoidal" data sets, and ≤ 4 % for "Paraboloid" data set).

However, the difference between computing times is distinct, the projection has been obtained by DMA 7 times quicker. This difference of computing time decreases by increasing the amount of vectors in the analysed data set and decreasing parameter k, because data preprocessing for iteration process, using Strategy II, requires more calculations.

Investigation of Relative MDS algorithm. RPM algorithm depends on various factors like strategies to select basic vectors, manner to initiate vectors in two-dimensional plane, the number of basic vectors. In this dissertation is presented and analyzed two new ways to select coordinates of vectors in two-dimensional plane of projection: the closest coordinates of basic vector or two input vector coordinates with the largest dispersion are selected.

Increasing the number of basic vectors is not always result in decreasing error. There are possibilities that error increase while number of basic vectors increase, and the selection of appropriate number of basic vectors guaranties less error in projection with relative MDS algorithm than best error, derived from SMACOF algorithm. The research was made with a view to determine more accurate way to select the number of vectors. It was found that in selection of number of basic vectors there is a limit, exceeding which error in most cases start to increase, if the number of basic vectors is increasing.

Problem of initialization of vectors. The errors of projection of various data sets are presented in (Table 2 and Table 3).

Table 2. Projection errors of Sammon's mapping

Dataset	The ways of initialization			
	Random	Line	PCA	By variance
"HBK"	0.006483	0.01140	**0.00464**	0.00555
"Wood"	0.025269	0.02536	**0.02432**	0.02537
"Iris"	0.004997	0.00491	**0.00397**	0.00406
"Wine"	0.000140	0.00012	**0.00003**	**0.00003**
"Cluster"	0.071625	0.07103	0.07115	**0.06667**

They are obtained using different initialization ways for Sammon's mapping (Table 2) and for the MDS SMACOF (Table 3). The experiments show that the smallest projection errors are obtained by using PCA, or by variance methods, but the PCA method is much more computing expensive. Therefore, we chose the variance method for the further experiments.

Table 3. Mean projection error ($\boldsymbol{E_{norm}}$) and computing time ($\boldsymbol{t}$) obtained by MDS SMACOF algorithm

Ways of initialization	Mean error, mean time	"Abalone"	"Paraboloid"	"Gaussian"	"Spheres"
Random	$\sqrt{E_{norm}}$	0.013019	0.209435	0.284020	0.219793
	t, s	233.81	**78.46**	90.80	25.20
Line	$\sqrt{E_{norm}}$	0.020931	0.209510	0.277183	0.219941
	t, s	**233.53**	78.50	**90.64**	25.20
By variance	$\sqrt{E_{norm}}$	0.013019	0.208405	0.273857	0.218949
	t, s	233.81	78.52	90.87	**25.12**
PCA	$\sqrt{E_{norm}}$	**0.012513**	**0.208306**	**0.272727**	**0.217274**
	t, s	234.48	79.17	91.80	25.50375

A dependence of the error on the way of initialization and the number of iterations are presented in Fig. 7. The results show that PCA and largest variances initialization ways are much better in the sense of the error than on the line, and slight better than random initialization of the vectors.

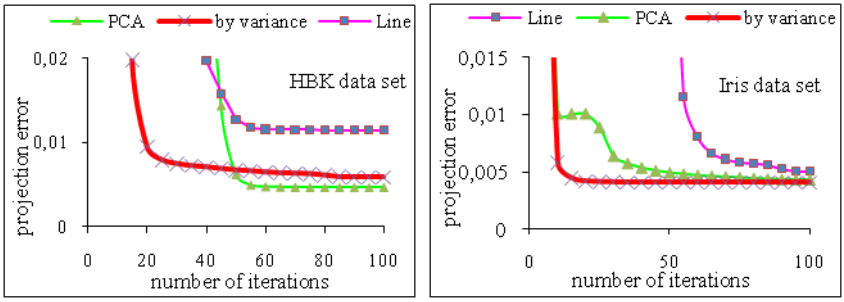

Fig. 7. Dependence of projection error on the way of initialization: "HBK" data set; "Iris" data set

In order to verify the obtained results, the analogical experiments have been carried out using larger data sets. Some additional multidimensional scaling algorithms (relative MDS and DMA) are also performed (Table 4). The obtained results show that the best ways of the initialization of the vectors are the largest variances and PCA.

Comparative analysis of some MDS algorithms. Three MDS algorithms (SMACOF, DMA and relative MDS) have been investigated and compared in order to answer a question which algorithm is suitable for visualization of large data sets. The algorithms have been examined using quantitative criteria of mapping.

Table 4. Quantitative criteria of mapping using multidimensional scaling algorithms

Criteria	SMACOF					
	"Spheres"	"Gaussian"	"Paraboloid"	"Elipsoidal"	"Abalone"	"Satimage"
MW	**4229.69**	**14844.34**	**115.6487**	184.6303	109.3453	67255.92
Spearmen coef.	**0.861522**	**0.812781**	**0.893843**	0.928474	0.999592	0.980790
Error	**0.217515**	**0.273772**	**0.208293**	0.207143	0.012816	0.1165890
Time	73.46	268.87	232.7953	360.58	693.79	2717.75
Criteria	Relative MDS with vectors initialization by variance					
MW	4237.45	14554.89	116.91	**179.42**	**108.01**	68713.81
Spearmen coef.	0.859215	0.810587	0.885582	**0.928548**	0.999592	0.981856
Error	0.219058	0.274714	0.213209	0.207212	0.012779	0.109482
Time	**35.78**	**38.29**	**37.33**	**39.85**	**42.43**	120.77
Criteria	Relative MDS with vectors initialization by PCA					
MW	4698.30	16594.15	121.04	188.57	108.53	**63134.47**
Spearmen coef.	0.811134	0.753909	0.878545	0.930524	**0.999597**	**0.985516**
Error	0.272489	0.301998	0.251414	**0.205049**	**0.012656**	**0.0952139**
Time	36.14	39.04	37.76	40.22	42.74	**95.59**
Criteria	DMA					
MW	4728.84	15310.89	229.78	248.8662	112.1474	266496.6285
Spearmen coef.	0.858383	0.811120	0.888252	0.927513	0.999583	0.915439
Error	0.219763	0.274438	0.212582	0.208014	0.012949	0.204888
Time	52.75	116.16	104.76	131.98	189.76	302.61

Often the algorithm is optimized according one criterion and it yield by other criteria. Sometimes all criteria are not equally important therefore weights of criteria are introduced in order to find the best solution of the pending problem. When large data sets are visualized, the first important criterion is the computing time, and the second one is the projection error. The relative MDS algorithm with initialization by variances is the best in five from six cases analyzed according the computing time. The relative MDS algorithm with initialization by PCA takes the second place, because the computing time is worse, but the projection error is smaller.

General Conclusions

Research, done in this work, revealed new possibilities of visualization methods grounded on multidimensional scaling.

Theoretical and experimental research led to the following conclusions:

1. There was proved theoretically that selection of initial points on the line in Sammon's mapping algorithm, when its slope coefficient is $a = \pm 1$, is not applicable. In theory, when this kind of initiation of points is applied, these points should stay on the same line. Because of computation and rounding errors, these points deviate from the line, and after several iterations disperse all around two-dimensional projection plane. Thus, it is advisable to use the following initiation ways: analysis of the main components or the largest dispersions method. Analysis of the main components and the largest dispersions method are much better in the sense of error that that of initiation on the line.
2. Comparison of the results, worked out using different multidimensional scaling-type algorithms, showed that the largest dispersion method is the best way to select the initial vectors in two-dimensional plane. This method makes the convergence of error quicker, and after several iterations the error is already sufficiently close to minimal projection error.
3. Visualizing large data sets and saving calculation time, it is effective to apply diagonal majorization algorithm (DMA). However, attention should be paid to the choice of strategy to order multidimensional vectors of set analyzed, and selection of parameter k of neighbourhood order. Examining the dependence of DMA algorithm results on multidimensional vectors order strategy showed smaller projection error regarding smaller number of neighbours k. All that allows reducing time of calculation up to three times, when $k \approx s/10$.
4. Diagonal majorization algorithm error is larger than SMACOF algorithm projection error, but when neighbourhood order parameter is set to

$k \geq 400$ or $k \approx s/10$ (for the analyzed data sets) and two different strategies to select neighbours are applied, the difference of those errors is less than 5 %. Selection of neighbours is performed here by changing the order numbers of vectors from the analysed data set at the beginning of DMA or after each iteration.

5. The number of retrained neurons in the SOM decreases and has a staircase form while the training epoch order number is increasing, and decrease by one after $e' = \left[\frac{n'e}{k'}\right] - \left[\frac{(n'-1)e}{k'}\right]$ $(n' = 1, \dots, k' - 2)$ epoch.
6. If the vectors from the analysed data set are not rationed according to the length, then it is possible to use coloring of the SOM cells in tones of grey that correspond to the length of neuron in the cell. In this representation, the position of neuron in the SOM net indicates the similarity of this neuron to other neurons in the sense of orientation of the codebook vector, and the color shows similarity in the sense of codebook vector length.
7. Combined algorithms SOM_Sammon and SOM_SMACOF assures similar quality of multidimensional data projection. This allows to apply not only often used combination of SOM and Sammon, but also combination of SOM and SMACOF algorithms, that is similar to the former and allows to save the computing time.
8. In relational perspective map algorithm, it is possible to use function of distance that allows the convergence of error minimization algorithm without using any additional parameters to stimulate the convergence.

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

Articles in the reviewed scientific periodical publications:

1. Bernatavičienė J., Dzemyda G., Marcinkevičius V. Conditions for Optimal Efficiency of Relative MDS, *Informatica*, 2007, Vol. 18(2), 187–202. ISSN 0868-4952. (*Current Abstracts. IAOR: International Abstracts In Operations Research. INSPEC. MatSciNet. ISI Web of Science. Scopus. TOC Premier. VINITI. Zentralblatt MATH*)

2. Bernatavičienė J., Dzemyda G., Marcinkevičius V. Diagonal Majorization Algorithm: Properties and Efficiency, *Information Technology and Control*, 2007, Vol. 36(4), 353–358. ISSN 1392-124X. (*ISI Web of Science. VINITI. INSPEC*)

3. Bernatavičienė J., Dzemyda G., Kurasova O., Marcinkevičius V. Optimal Decisions in Combining the SOM with Nonlinear Projection Methods, *European Journal of Operational Research*, Elsevier, 2006, Vol. 173(3),

729–745. ISSN 0377-2217. (*ISI Web of Science. Science Direct. INSPEC. Business Source Complete. GeoRef. Computer Abstracts International Database. Compendex*)

4. Bernatavičienė J., Dzemyda G., Kurasova O., Marcinkevičius V. Strategies of Selecting the Basic Vector Set in the Relative MDS, *Technological and Economic Development of Economy*, 2006, Vol. 12(4), 283–288. ISSN 1392-8619. (*ASCE Civil Engineering Abstracts. Business Source Complete. Business Source Premier. Current Abstracts. ICONDA. SCOPUS. TOC Premier*)

5. Karbauskaitė R., Marcinkevičius V., Dzemyda G. Testing the Relational Perspective Map for Visualization of Multidimensional Data, *Technological and Economic Development of Economy*, 2006, Vol. 12(4), 289–294. ISSN 1392-8619. (*ASCE Civil Engineering Abstracts. Business Source Complete. Business Source Premier. Current Abstracts. ICONDA. SCOPUS. TOC Premier)*

6. Dzemyda G., Bernatavičienė J., Kurasova O., Marcinkevičius V. Strategies of Minimization of Sammon's Mapping Error. *Lithuanian Mathematical Journal*, 2004, Vol. 44, Spec. no., 1–6. ISSN 0132-2818, in Lithuanian. (*MatSciNet. CIS: current index to statistics. VINITI. Zentralblatt MATH*)

7. Dzemyda G., Kurasova O., Marcinkevičius V. Parallelization in Combining the SOM and Sammon's Mapping. *Lithuanian Mathematical Journal*, 2003, Vol. 43, Spec. no., 218–222. ISSN 0132-2818, in Lithuanian. *(MatSciNet. CIS: current index to statistics. VINITI. Zentralblatt MATH*)

8. Dzemyda G., Kurasova O., Marcinkevičius V. Application of MPI Software Package in Parallel Visualization. *Information Sciences*, 2003, Vilnius, Vilniaus universitetas, No. 26, 230–235. ISSN 1392-0561, in Lithuanian.

Articles in the other editions:

9. Karbauskaitė R, Dzemyda G., Marcinkevičius V. Selecting a Regularisation Parameter in the Locally Linear Embedding Algorithm, *The 20th International Conference EURO Mini Conference "Continuous Optimization and Knowledge-Based Technologies" EurOPT'2008: May 20-23*, 2008, Neringa, Lithuania: selected papers. Vilnius: Technica, 59–64. *(Conference Proceedings Citation Index*)

10. Marcinkevičius V. Statistical Estimation of the Multidimensional Data Visualization Algorithms, *Science and Supercomputing in Europe Report 2007*, 2008, Bologna: CINECA *Consorzio Interuniversitario*, 382–384. ISBN 978-88-86037-21-1.

11. Bernatavičienė J., Dzemyda G., Kurasova O., Marcinkevičius V., Medvedev V. The Problem of Visual Analysis of Multidimensional Medical Data. Models and Algorithms for Global Optimization, *Springer Optimization and Its Applications*, 2007, New York, Springer, Vol. 4, 277–298. ISBN 978-0-387-36720-9. (*SpringerLINK*)

12. Bernataviciene, J., Dzemyda, G., Kurasova, O., Marcinkevičius, V. Decision Support for Preliminary Medical Diagnosis Integrating the Data Mining Methods, *Simulation and Optimisation in Business and Industry: 5th International Conference on Operational Research*: May 17–20, 2006, Kaunas, Technologija, 155–160. ISBN 9955-25-061-5. (*ISI Proceedings*)

13. Dzemyda G., Bernatavičienė J., Kurasova O., Marcinkevičius V. Minimization of the Mapping Error Using Coordinate Descent, *The 13-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2005 in Co-operation with Eurographics*, 2005, Plzen, University of West Bohemia, 169–172. ISBN 80-903100-9-5.

14. Marcinkevičius V., Dzemyda G. Visualization of the Multidimensional Data Using the Trained Combination of SOM and Sammon's Algorithm, *Information Technologies 2004 the Conference Materials*, 2004, Kaunas, Technologija, 350–355. ISBN 9955-09-588-1.

About the author

Virginijus Marcinkevičius was born in Alytus on the 21th of June in 1976. After finishing the Alytus "Piliakalnio" secondary school in 1994, he graduated from Vilnius Pedagogical University in 2001 and acquired a bachelor's degree in mathematics and informatics. In 2003 he acquired a master's degree in mathematics. Since 2001 he is employee of the Institute of Mathematics and Informatics. From 2003 till 2008 he has been a PhD student at same institutions. He is a member of the Computer Society and Lithuanian Mathematical Society.

E-mail: VirgisM@ktl.mii.lt.

NETIESINĖS DAUGIAMAČIŲ DUOMENŲ PROJEKCIJOS METODŲ SAVYBIŲ TYRIMAS IR FUNKCIONALUMO GERINIMAS

Mokslo problemos aktualumas

Duomenų suvokimas yra sudėtingas procesas, ypač kai duomenys nurodo sudėtingą objektą, reiškinį, kuris apibūdinamas daugeliu kiekybinių ir kokybinių parametrų ar savybių. Tokie duomenys vadinami daugiamačiais duomenimis ir gali būti interpretuojami kaip taškai arba vietos vektoriai daugiamatėje erdvėje. Analizuojant daugiamačius duomenis, dažnai į pagalbą pasitelkiame vieną svarbiausių duomenų analizės įrankių – duomenų vizualizavimą arba grafinį informacijos pateikimą. Pagrindinė vizualizavimo idėja – duomenis pateikti tokia forma, kuri leistų naudotojui lengviau suprasti duomenis, daryti išvadas ir tiesiogiai įtakoti tolesnį sprendimų priėmimo procesą. Vizualizavimas leidžia geriau suvokti sudėtingas duomenų aibes, gali padėti nustatyti tyrėją dominančius jų poaibius. Dimensijos mažinimo metodai leidžia atsisakyti tarpusavyje priklausomų duomenų komponenčių, o projekcijos metodais galima transformuoti daugiamačius duomenis į tiesę, plokštumą, trimatę erdvę ar į kitą žmogui vizualiai suvokiamą formą. Vizualią informaciją žmogus pajėgus suvokti daug greičiau ir paprasčiau negu skaitinę arba tekstinę. Iš kitos pusės toks suvokimas gali būti tik kaip dirva hipotezėms ir tolimesniems tyrimams, pagrįstiems griežtais matematiniais modeliais. Kokia informacija ir kaip ji turi būti vizualiai pateikiama, priklauso nuo naudotojo, dirbančio šioje srityje, todėl čia iškyla problemos, reikalaujančios atsakymų: kokius vizualizavimo metodus pasirinkti ir kaip optimaliai parinkti jų parametrus. Dėl nuolat didėjančių duomenų aibių, atsiranda vis nauji duomenų vizualizavimo metodai, tačiau išlieka aktuali problema – šių metodų aprobavimas ir taikymo pagrįstumo tyrimai.

Tyrimo objektas

Disertacijos tyrimo objektas yra daugiamačiai duomenys, jų atvaizdavimas netiesiniais daugiamačių skalių algoritmais ir saviorganizuojančiais neuroniniais tinklais, projekcijos kokybės vertinimas.

Darbo tikslas ir uždaviniai

Darbo tikslas – netiesinės daugiamačių duomenų projekcijos metodų funkcionalumo gerinimas, tiriant jų savybes.

Siekiant šio tikslo sprendžiami šie uždaviniai:

1. Ištirti daugiamačių skalių algoritmų duomenų pradinio parinkimo būdus.
2. Palyginti daugiamačių skalių SMACOF algoritmą, Sammono algoritmą ir santykinių daugiamačių skalių algoritmą topologijos išsaugojimą įvertinančiais kriterijais.
3. Ištirti diagonalinio mažoravimo algoritmo efektyvumą, lyginant jį su daugiamačių skalių SMACOF realizacija ir santykinių daugiamačių skalių algoritmais.
4. Teoriškai ištirti saviorganizuojančio neuroninio tinklo (SOM) neuronų nugalėtojų skaitinę priklausomybę nuo mokymo epochos.
5. Ištirti naujas galimybes SOM tinklui vaizduoti.
6. Modifikuoti santykinės perspektyvos metodo algoritmą, siekiant pagerinti jo konvergavimą.
7. Ištirti santykinių daugiamačių skalių algoritmo parametrus, siekiant apskaičiuoti vienareikšmišką ir tikslią projekciją.

Tyrimų metodika

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

Kuriant programinę įrangą naudotas programinio modeliavimo metodas. Teoriniai tyrimo metodai naudoti įrodant teoremas ir tiriant algoritmų konvergavimą. Taikytas matematinės indukcijos principas įrodant teiginius.

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

Mokslinis naujumas

Darbe atlikti tyrimai atskleidė naujas galimybes vystyti daugiamačių duomenų vizualizavimo metodus ir priemones.

Įrodyta, kad Sammono algoritme projekcijos duomenų pradinis parinkimas ant tiesės yra netinkamas. Remiantis tuo yra tikslinga naudotis

principinių komponenčių analize ar didžiausių dispersijų metodu parenkant projekcijos pradinius taškus.

Parodyta, kad diagonalinio mažoravimo algoritmo efektyvumas nusileidžia daugiamačių skalių SMACOF realizacijai ir santykinėms daugiamatėms skalėms.

Teoriškai ištirta vienos epochos metu perskaičiuojamų stačiakampės formos SOM tinklo neuronų skaičiaus priklausomybė nuo mokymo epochos numerio.

Pasiūlytas naujas būdas neuroninio tinklo SOM vaizdavimui. Jame neuroninio tinklo lentelės ląstelių spalva parenkama kaip pilkos spalvos atspalvis, priklausantis nuo ląstelėse esantį neuroną atitinkančio vektoriaus ilgio.

Pasiūlytas naujas pradinių duomenų parinkimo būdas pagal didžiausias dispersijas, tinkamas visiems daugiamačių skalių klasės algoritmams.

Ištirtas santykinės perspektyvos metodo konvergavimas, ir pasiūlyta naudoti dvi naujas atstumų funkcijas, taip užtikrinat RPM metodo konvergavimą.

Praktinė vertė

Tyrimų rezultatai taikyti tiriamuosiuose Lietuvos valstybinio mokslo ir studijų fondo ir Lietuvos mokslo tarybos projektuose:

- Prioritetinių Lietuvos mokslinių tyrimų ir eksperimentinės plėtros programoje „Informacinės technologijos žmogaus sveikatai – klinikinių sprendimų palaikymas (e-sveikata), IT sveikata“; registracijos Nr.: C-03013; vykdymo laikas: 2003 m. 09 mėn. – 2006 m. 10 mėn.
- Aukštųjų technologijų plėtros programos projekte „Žmogaus genomo įvairovės ypatumų nulemti aterosklerozės patogenezės ypatumai (AHTHEROGEN)“; registracijos Nr.: U-04002; vykdymo laikas: 2004 m. 04 mėn. – 2006 m. 12 mėn.
- Aukštųjų technologijų plėtros programos projekte „Informacinės klinikinių sprendimų palaikymo ir gyventojų sveikatinimo priemonės e. Sveikatos sistemai (Info Sveikata)“; registracijos Nr.: B-07019; vykdymo laikas: 2007 m. 09 mėn. – 2009 m. 12 mėn.
- Prioritetinių Lietuvos mokslinių tyrimų ir eksperimentinės plėtros krypčių projekte „Genetinių ir genominių lūpos ir (arba) gomurio nesuaugimo pagrindų tyrimai (GENOLOG)“; registracijos Nr.: C-07022; vykdymo laikas: 2007 m. 04 mėn. – 2009 m. 12 mėn.
- Dvišalio bendradarbiavimo mokslo tyrimų ir eksperimentinės plėtros srityje Lietuvos – Prancūzijos integruotos veiklos programoje „Žiliberas“;

registracijos Nr.: V-09059; vykdymo laikas: 2008 m. 04 mėn. – 2010 m. 12 mėn.

Ginamieji teiginiai

1. Sammono algoritme projekcijos duomenų iniciacija ant tiesės yra netinkama, kadangi paklaidos konvergavimas iteracinio proceso pradžioje yra lėtas.
2. Diagonalinis mažoravimo algoritmas paklaidos prasme nusileidžia daugiamačių skalių SMACOF algoritmui ir santykinėms daugiamatėms skalėms. DMA paklaida gaunama didesnė už SMACOF ir santykinių daugiamačių skalių algoritmo paklaidas, tačiau DMA yra greitesnis už SMACOF algoritmą.
3. Stačiakampės formos SOM tinklo, kurio didesniąją briauną sudarančių neuronų yra k', permokomų neuronų skaičius laiptiškai mažėja didėjant mokymo epochos eilės numeriui ir sumažėja vienetu po $e' = \left[\frac{n'e}{k'}\right] - \left[\frac{(n'-1)e}{k'}\right]$ $(n' = 1, \dots, k' - 2)$ epochos.
4. Galimos naujos atstumų funkcijos, kurios žymiai pagerina RPM algoritmo veikimą.
5. Pradinių taškų parinkimo pagal didžiausias dispersijas būdas, daugiamačių skalių algoritmuose yra vienas tiksliausių ir efektyviausių.

Darbo apimtis

Disertaciją sudaro įvadas, trys skyriai ir rezultatų apibendrinimas. Darbo apimtis yra 105 puslapiai, neskaitant priedų, tekste panaudotos 57 numeruotos formulės, 29 paveikslai ir 13 lentelių. Rašant disertaciją buvo panaudotas 107 literatūros šaltinis.

Bendrosios išvados

1. Teoriškai įrodyta, kad Sammono projekcijos algoritme pradinių taškų parinkimas ant tiesės, kai jos krypties koeficientas lygus $a = \pm 1$, yra netaikytinas. Teoriškai, naudojant tokią taškų iniciaciją, šie taškai turėtų išlikti ant tos pačios tiesės. Dėl skaičiavimo ir skaičių apvalinimo paklaidų taškai palieka tiesę ir po keleto iteracijų išsibarsto po visą dvimatę projekcijos plokštumą. Tikslinga naudoti tokius iniciacijos būdus, kaip pagrindinių komponenčių analizė ar didžiausių dispersijų metodas. Pagrindinių komponenčių analizės ir didžiausių dispersijų iniciacijos metodai yra žymiai geresni paklaidos prasme už iniciaciją ant tiesės.

2. Palyginus rezultatus, gautus naudojant skirtingus daugiamačių skalių tipo algoritmus, nustatyta, kad optimalu pradinius vektorius dvimatėje plokštumoje parinkti naudojant didžiausių dispersijų metodą. Šis iniciacijos metodas pagreitina paklaidos konvergavimą ir jau po pirmųjų iteracijų gaunama pakankamai artima minimaliai projekcijos paklaida.
3. Tyrimai parodė, kad vizualizuojant dideles duomenų aibes ir taupant skaičiavimo laiką, efektyvu naudoti diagonalinį mažoravimo algoritmą. Tačiau reikia atkreipti dėmesį į analizuojamos aibės daugiamačių vektorių rikiavimo strategijos ir kaimyniškumo eilės parametro k parinkimą. Ištyrus DMA algoritmo rezultatų priklausomybę nuo daugiamačių vektorių rikiavimo strategijos, gautos mažesnės projekcijos paklaidos atsižvelgiant į mažesnį kaimynų skaičių k. Visa tai leidžia iki trijų kartų sutaupyti skaičiavimo laiką, kai $k \approx s/10$.
4. Diagonalinio mažoravimo algoritmo projekcijos paklaida yra didesnė už SMACOF algoritmo projekcijos paklaidą, tačiau, parenkant kaimyniškumo eilės parametrą $k \geq 400$ arba $k \approx s/10$ (tirtoms aibėms) ir naudojant kaimynų perrikiavimo strategijas, kai kaimynai perrikiuojami algoritmo pradžioje arba po kiekvienos iteracijos, šių paklaidų skirtumas yra mažesnis už 5.
5. SOM tinklo permokomų neuronų skaičius laiptiškai mažėja didėjant mokymo epochos eilės numeriui ir sumažėja vienetu po $e' = \left[\frac{n'e}{k'}\right] - \left[\frac{(n'-1)e}{k'}\right]$ $(n' = 1, \dots, k' - 2)$ epochos.
6. Jeigu analizuojamos duomenų aibės vektoriai nėra sunormuoti pagal vektoriaus ilgį, tuomet galima naudoti SOM tinklo ląstelių spalvinimą pilkos spalvos atspalviais, priklausančiais nuo ląstelės neurono ilgio. Šiame vaizdavime neurono padėtis SOM tinkle nurodo jo panašumą į kitus tinklo neuronus pagal kryptį, o spalva – pagal neurono ilgį.
7. SOM_Sammono ir SOM_SMACOF junginiai yra užtikrina panašią daugiamačių duomenų projekcijos kokybę. Tai leidžia taikyti ne tik dažai naudojamą SOM ir Sammono junginį, bet ir jam panašų SOM ir SMACOF algoritmų junginį, taip sutaupant skaičiavimas reikalingo laiko.
8. RPM algoritme galima naudoti atstumų funkciją, leidžiančią paklaidos minimizavimo algoritmui konverguoti nenaudojant papildomų konvergavimą skatinančių parametrų.

Virginijus MARCINKEVIČIUS

NETIESINĖS DAUGIAMAČIŲ DUOMENŲ
PROJEKCIJOS METODŲ SAVYBIŲ
TYRIMAS IR FUNKCIONALUMO GERINIMAS

Daktaro disertacija

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

Virginijus MARCINKEVIČIUS

INVESTIGATION AND FUNCTIONALITY IMPROVEMENT OF NONLINEAR
MULTIDIMENSIONAL DATA PROJECTION METHODS

Doctoral Dissertation

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

2010 08 20 . 1 sp. l. Tiražas 60 egz.
Išleido Matematikos ir informatikos institutas
Akademijos g. 4, LT-08663 Vilnius.
Interneto svetainė: *http://www.mii.lt.*
Spausdino „Kauno technologijos universiteto spaustuvė“,
Studentų g.54, LT-51424 Kaunas