VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Viktor MEDVEDEV

RESEARCH OF MULTIDIMENSIONAL DATA VISUALIZATION USING FEED-FORWARD NEURAL NETWORKS

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

VGTU
LEIDYKLA
Vilnius TECHNIKA 2007

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2003–2007.

Scientific Supervisor

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

The dissertation is being defended at the Council of Scientific Field of Informatics Engineering at Vilnius Gediminas Technical University:

Chairman:

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

Members:

Prof Dr Habil Feliksas IVANAUSKAS (Vilnius University, Physical Sciences, Informatics – 09P),

Assoc Prof Dr Regina KULVIETIENĖ (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

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

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

Opponents:

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

Assoc Prof Dr Antanas Leonas LIPEIKA (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Informatics Engineering in the Conference and Seminars Center of the Institute of Mathematics and Informatics at 11 a. m. on January 17 2008.

Address: Goštauto str. 12, LT-01108 Vilnius, Lithuania.

Tel.: +370 5 274 4952, +370 5 274 4956; fax +370 5 270 0112;

e-mail: doktor@adm.vgtu.lt

The summary of the doctoral dissertation was distributed on 17 December 2007.

A copy of the doctoral dissertation is available for review at the Library of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and at the Library of Institute of Mathematics and Informatics (Akademijos g. 4, LT-08663 Vilnius, Lithuania)

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Viktor MEDVEDEV

TIESIOGINIO SKLIDIMO NEURONINIŲ TINKLŲ TAIKYMO DAUGIAMAČIAMS DUOMENIMS VIZUALIZUOTI TYRIMAI

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

Vilnius LEIDYKLA TECHNIKA 2007

Disertacija rengta 2003–2007 metais Matematikos ir informatikos institute.

Mokslinis vadovas

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

Disertacija ginama Vilniaus Gedimino technikos universiteto Informatikos inžinerijos mokslo krypties taryboje:

Pirmininkas:

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

Nariai:

prof. habil. dr. Feliksas IVANAUSKAS (Vilniaus universitetas, fiziniai mokslai, informatika – 09P),

doc. dr. Regina KULVIETIENĖ (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

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

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

Oponentai:

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

doc. dr. Antanas Leonas LIPEIKA (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija bus ginama viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2008 m. sausio mėn. 17 d. 11 val. Matematikos ir informatikos instituto konferencijų ir seminarų centre.

Adresas: Goštauto g. 12, LT-01108 Vilnius, Lietuva.

Tel.: (8 5) 274 4952, (8 5) 274 4956; faksas (8 5) 270 0112;

el. paštas doktor@adm.vgtu.lt

Disertacijos santrauka išsiuntinėta 2007 m. gruodžio 17 d.

Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) ir Matematikos ir informatikos instituto (Akademijos g. 4, LT-08663 Vilnius, Lietuva) bibliotekose.

VGTU leidyklos „Technika“ 1439 mokslo literatūros knyga.

General Characteristic of the Dissertation

Topicality of the problem. The research area of this work is the analysis of multidimensional data and the ways of improving apprehension of the data. Data apprehension is rather a complicated problem especially if the data refer to a complex object or phenomenon described by many parameters. A tendency has been recently observed that scientists, who pursue investigations of MDS (multidimensional scaling), frequently dissociate from other methods of research or even ignore them. On the other hand, in other investigations of visualization methods there are no comparisons or connections with MDS-type methods. In this work, we try to extend the realizations of MDS-type methods by applying artificial neural networks, thus strengthening the relationship among different trends of visual data analysis. The work deals with artificial neural network algorithms for visualizing multidimensional data. A specific learning rule (SAMANN) has been proposed. That allows a feed-forward neural network to realise Sammon's projection.

Minimization of the projection error of multidimensional data by using artificial neural networks is the main **problem** addressed in this dissertation.

Aim and tasks of the work. The key aim of the work is to develop and improve methods how to efficiently minimize visualization errors of multidimensional data by using artificial neural networks. It was necessary to solve these tasks: 1) to analyse the methods of multidimensional data visualization; 2) to investigate the abilities of artificial neural networks to visualize multidimensional data; 3) to create parallel realizations of the SAMANN algorithm; 4) to improve and speed-up the training and retraining process of the SAMANN algorithm; 5) to search for the optimal values of the algorithm learning rate; 6) to investigate the abilities of the artificial neural networks in projecting new data.

Research object. The research object of the dissertation are artificial neural networks for multidimensional data projection. General topics related with this object are: 1) multidimensional data visualization; 2) dimensionality reduction algorithms; 3) errors of projecting data; 4) projection of the new data; 5) strategies for retraining the neural network that visualizes multidimensional data; 6) optimization of control parameters of the neural network for multidimensional data projection; 7) parallel computing.

Scientific novelty. A parallel realization of the SAMANN algorithm for multidimensional data projection has been created. The strategies for retraining

the neural network have been proposed. It has been established experimentally how to select the learning parameter value of the SAMANN neural network so that the algorithm would work efficiently.

The research **methodology** is based on the development of new strategies for SAMANN neural network training and their experimental investigations.

Practical value. The results of the research are applied in solving some problems in practice. Human physiological data that describe the human functional state have been investigated. The results, obtained by the method, can be of use to medics for a preliminary diagnosis: healthy, unclear, or sick persons.

The base of research of psychological data is the project "Information technologies for human health – clinical decision support (e-Health). IT Health (No. C-03013)", supported by the Lithuanian State Science and Studies Foundation.

Approbation and publications of the research. The main results of this dissertation were published in 11 scientific papers: 2 articles in periodical scientific publications from the ISI Web of Science list; 2 articles in periodical scientific publications from the ISI Proceedings list; 1 article in the book published by Springer; 1 chapter of the book published by IOS Press; 3 articles in periodical scientific publications from the list approved by the Science Council of Lithuania; 2 articles in the proceedings of scientific conferences. The main results of the work have been presented and discussed at 4 international and 5 national conferences.

The scope of the scientific work. The work is written in Lithuanian. It consists of 9 chapters, and the list of references. There are 144 pages of the text, 86 figures, 1 table and 159 bibliographical sources.

1. Introduction

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

2. Analysis of the Methods of Multidimensional Data Visualization

The chapter is devoted to the review and analysis of the various methods of visualization. Multidimensional data, meaning the data that require more than two or three dimensions to represent, can be difficult to interpret. Direct

visualization methods and projection (dimensionality reduction) methods are investigated. Several approaches have been developed for visualizing high-dimensional data. Many of the methods, such as parallel coordinates, star glyphs, Chernoff faces, try to show all dimensions of the data at the same time. This approach is only suitable for relatively few dimensions. When the dimensionality of the data increases, some other means have to be used. One of the main strategies used to handle very high dimensional data is dimensionality reduction where the task is to reduce the dimensionality of the data to two or three for visualization. A large number of different projection methods have been developed for this task. There are linear and nonlinear projection methods. The principal component analysis, projection pursuit are linear projection methods; multidimensional scaling, principal curves, triangulation, isomap are nonlinear ones. A more precise data structure is preserved using nonlinear projection methods. Nevertheless, the projection errors are inevitable. It is necessary to look for the ways of minimizing these projection errors.

One of multidimensional scaling methods to map a high-dimensional space onto a space of lower dimensionality is Sammon mapping. Suppose that we have m data points, $X_i = (x_{i1}, x_{i2} \ldots, x_{im})$, $i = 1, \ldots, n$, in a n-space and, respectively, we define m points, $Y_i = (y_{i1}, y_{i2} \ldots, y_{im})$, $i = 1, \ldots, d$, in a d-space ($d<n$). The pending problem is to visualize these n-dimensional vectors $X_i, i = 1, ..., n$ onto the plane R^2. Let d_{ij}^* denote the distance between X_i and X_j in the input space, and d_{ij} denote the distance between the corresponding points Y_i and Y_j in the projected space. The Euclidean distance is frequently used. The projection error measure E (so-called Sammon's stress) is as follows: $E = \frac{1}{\sum_{i,j=1;i<j}^{m} d_{ij}^*} \sum_{\substack{i,j=1 \\ i<j}}^{m} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*}$. It is a measure of how well the distances are preserved when the patterns are projected to a lower-dimensional space.

3. Concepts of Artificial Neural Networks

Models of an artificial neuron, network architectures, some learning rules, self-organizing maps, radial basis function networks are described. Artificial Neural Networks (ANN) are algorithms inspired by biology. The idea is to build systems that reproduce the structure and functioning of the brain neurons. Research in this field began in the 1940s, with the works of McCullogh and Pitts, followed by

Hebb, Rosenblatt, Widrow. An Artificial Neural Network can be described as a set of interconnected adaptive units generally organized in a layered structure.

4. Artificial Neural Networks for Multidimensional Data Visualization

This chapter deals with the capabilities of ANN to visualize the multidimensional data. Application of artificial neural networks to Sammon's projection is analysed: a feed-forward neural network is trained by a specific back-propagation learning algorithm. The self-organizing neural networks application areas, curvilinear component analysis, autocoders, NeuroScale methods are discussed in this chapter.

In this work, an unsupervised backpropagation algorithm for training a multilayer feed-forward neural network (SAMANN) for perform Sammon's nonlinear projection is investigated. This algorithm preserves all interpattern distances as well as possible. Sammon mapping has a drawback. It lacks generalization, which means that new points cannot be added to the map obtained without recalculating it. The SAMANN network offers the generalization ability of projecting new data, which are not present in the original Sammon projection algorithm. It is a feed-forward neural network where the number of input units is set to be the feature space dimension n, and the number of output units is specified as the extracted feature space dimension d (Fig 1). Mao and Jain (J.Mao, A.K.Jain, Artificial neural networks for feature extraction and multivariate data projection, *IEEE Transactions on Neural Networks*, Vol. 6, No. 2, 1995, p. 296–317) have derived a weight updating rule for the multilayer perceptron neural network that minimizes Sammon's stress using the gradient descent method.

The SAMANN unsupervised backpropagation algorithm is as follows:

1. Initialize the weights randomly in the SAMANN network.

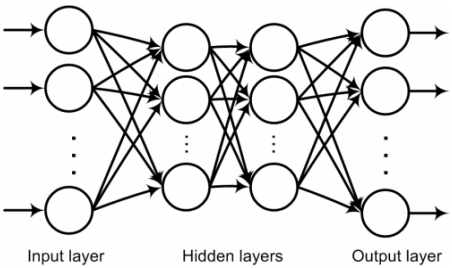

Fig 1. Feed-forward network for Sammon's projection (SAMANN)

2. Select a pair of patterns randomly, present them to the network one at a time, and evaluate the network in a feed-forward fashion.
3. Update the weights in the backpropagation fashion starting from the output layer.
4. Repeat steps 2–3 a number of times.
5. Present all the patterns and evaluate the outputs of the network; compute Sammon's stress; if the value of Sammon's stress is below a predefined threshold or the number of iterations (from steps 2–5) exceeds the predefined maximum number, then stop; otherwise, go to step 2.

5. Learning Problems of the SAMANN Neural Network

The rate, at which artificial neural networks learn, depends upon several controllable factors. When projecting data, it is of great importance to achieve good results in a short time interval. In the consideration of the SAMANN network, it has been observed that the projection error depends on different parameters. Investigations have revealed that, in order to achieve good results, one needs to correctly select the learning rate η. It has been stated so far that projection yields the best results if the η value is taken from the interval (0;1). In that case, the network training is very slow. One of the possible reasons is that, in the case of the SAMANN network, the interval (0;1) is not the best one. Thus, it is reasonable to look for the optimal value of the learning parameter that may not necessarily be within the interval. The experiments, done in this chapter, show in what way the SAMANN network training depends on the learning rate.

The experiments have been done with real and artificial datasets. At first the dependence of the data projection accuracy on the learning rate η has been defined for $\eta \in (0;1)$. The results obtained are illustrated in Fig 2. This figure demonstrates that with an increase in the learning rate value, a better projection error is obtained. That is why the experiments have been done with higher values of the learning rate beyond the limits of the interval (0;1). The results are presented in Fig 3. It has been noticed that the best results are at $\eta > 1$.

We can conclude from Figures 2 and 3 that the optimal value of the learning rate for the datasets considered is within the interval [10;30]. In the case of the Salinity dataset, the optimal value of the learning rate is $\eta = 10$, for the Iris dataset $\eta = 30$. At these values of the learning rate we obtain the best projection results, i.e., the data are projected more rapidly and more exactly. For the fixed number of iterations, good projection results are obtained in a shorter time interval than that taking the η values from the interval (0;1).

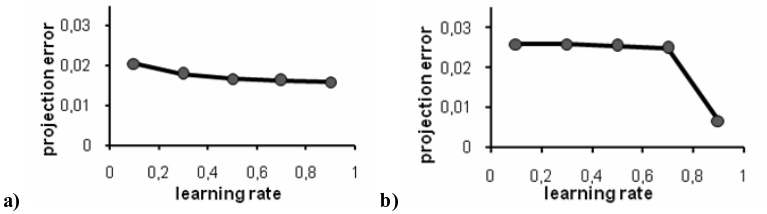

Fig 2. (a – Salinity dataset, b – Iris dataset) Dependence of the data projection accuracy on the learning rate η, $\eta \in (0,1)$

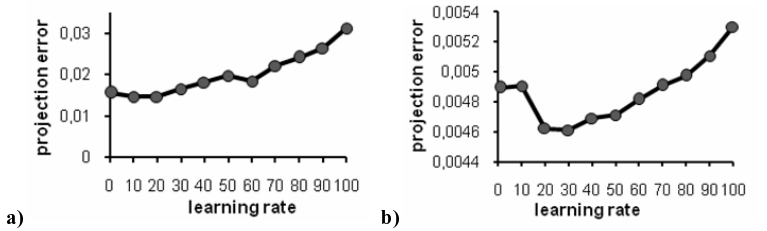

Fig 3. (a – Salinity dataset, b – Iris dataset) Dependence of the data projection accuracy on the learning rate η, $\eta \in [1,100)$

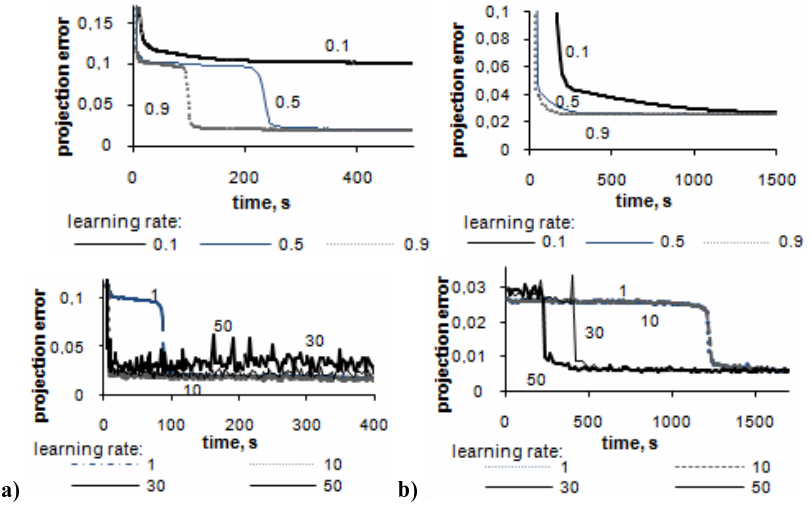

Fig 4. The dependence of the projection error on the computation time for the salinity dataset (a) and the iris dataset (b)

While experimenting the computing time and the errors obtained in each iteration have been defined as well. In Fig 4, the dependence of the projection error on the computation time is presented at different values of the learning rate for two real datasets (salinity and iris datasets). Fig 4 indicate that the higher the value of the learning rate, the more rapidly one succeeds in getting good results (i.e., sufficiently low projection error). However, with an increase in the value of the learning rate, the error variations also increase, which can cause certain network training problems. Fig 5 illustrates 2D projection maps of the datasets using the SAMANN network at the optimal value η of the learning rate, defined before.

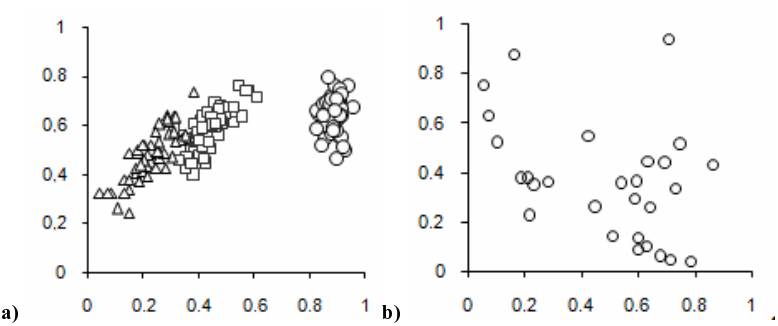

Fig 5. (a – Iris, b – Salinity dataset) 2D projections of datasets using the SAMANN

6. Parallel Realizations of the SAMANN Algorithm

A drawback of using SAMANN is that the training process is extremely slow. One of the ways of speeding up the network training process is to use parallel computing. In this chapter, some parallel realizations of the SAMANN are proposed. The training process takes a considerable amount of time. One of the ways of solving this problem and decreasing the computation time is to use parallel computing and adapt the methods of making the consecutive algorithms parallel (which allow us to use several processors for the net training at the same time).

Some strategies of the SAMANN parallel realization have been suggested and examined. One of the strategies is presented below. At the beginning of each iteration, the processor (p_0) randomly mixes the starting data array A and divides it into two parts A_1 and A_2, then p_0 distributes it among the rest two processors p_1 and p_2 (the starting data array elements are divided into equal parts). Two identical neural networks are created. Each network is trained with

an appropriate part of data array A_1 or A_2, using M_1 iterations. This process is fulfilled by two processors in parallel, the first processor uses array A_1 and the second one array A_2. Then the calculation is fulfilled by one processor. After the training (M_1 iter.), weight sets W_1 and W_2 of each net, output vectors and appropriate projection errors E_1 and E_2 have been got for the whole data array A. Projection errors E_1 and E_2 are obtained when all the vectors from array A are used in the nets trained appropriately with data arrays A_1 and A_2. Of two weight sets there remains only the one with a smaller projection error. Using the obtained weight set, the neural network is trained with all the data set A vectors (M_2 iterations).

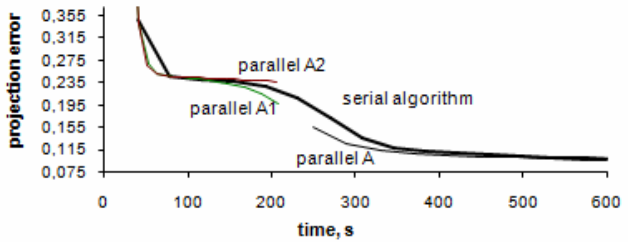

Fig 6. Dependence of the projection error on the computation time for the Iris dataset

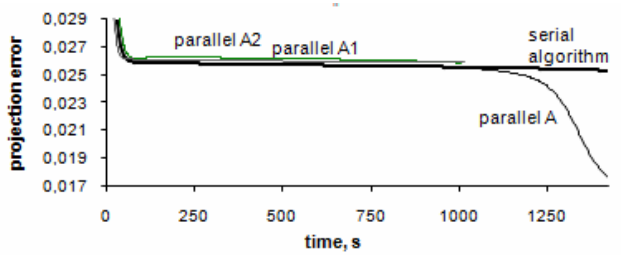

Fig 7. Dependence of the projection error on the computation time for the Ionosphere dataset

Two datasets were used in the experiments: Iris Dataset (150 4-dimensional vectors); Ionosphere Dataset (351 34-dimensional vectors). The mapping errors were calculated and the calculation time was measured for serial and parallel algorithms. The results of the parallel algorithm are compared with that of the serial one. The results for the Iris dataset and the Ionosphere dataset are presented in Figures 6 and 7. The figures show that a modified parallel algorithm makes it possible to achieve good visualization

results during a shorter time. The starting data set can be divided into k parts ($k \geq 2$). Good results were also achieved dividing the data set into 3 parts.

7. Retraining of the SAMANN Network

After training the SAMANN network, a set of weights of the neural network is fixed. A new vector, shown to the network, is mapped into the plane very fast and quite exactly without any additional calculations. However, while working with large data amounts, there may appear a lot of new vectors, which entails retraining of the SAMANN network after some time. That is why strategies for retraining the neural network that realizes multidimensional data visualization have been proposed and then their analysis made. Retraining of the network has to be efficient and the training algorithm has to converge rapidly.

The strategies of the neural network data retraining are as follows (two strategies are presented below):

1. The SAMANN network is trained by N_1 initial vectors, a set of weights w_1 is obtained, then the visualization error $E(N_1)$ is calculated, and vector projections are localized on the plane. After the emergence of N_2 new vectors, the neural network is retrained with all the N_1+N_2 vectors, and after each iteration the visualization error $E(N_1+N_2)$ is calculated, the computing time is measured. A new set of the network weights w_2 is found.
2. The SAMANN network is trained by N_1 initial vectors, a set of weights w_1 is obtained, and the visualization error $E(N_1)$ is calculated. Since in order to renew the weights, a pair of vectors μ and ν is simultaneously provided for the neural network, the neural network is retrained with $2*N_2$ vectors at each iteration: at each training step one vector is taken from the primary dataset and the other from the new one. After each iteration the visualization error $E(N_1+N_2)$ is calculated and the computing time is measured. A new set of network weights w_2 is found.

Two datasets have been used in the experiments: Iris or Fisher's Dataset (a real dataset with 150 random samples of iris flowers, described by 4 attributes); 300 randomly generated vectors (three spherical clusters with 100 5-dimensional vectors each). In the process of calculating, the time of algorithm performance was measured. Figures 8 and 9 demonstrate the results of calculations. Only the results of retraining the SAMANN network with the new vectors are indicated in the figures. The best visualization results are obtained by taking points for network retraining from the primary dataset and the new dataset (second strategy). The second strategy enables us to attain good

visualization results in a very short time as well as to get smaller visualization errors and to improve the accuracy of projection as compared to other strategies (Fig 9 illustrates this fact best). The second strategy proposed makes it possible to reduce the duration of computing a great deal in case there are considerably less new vectors than the initial ones.

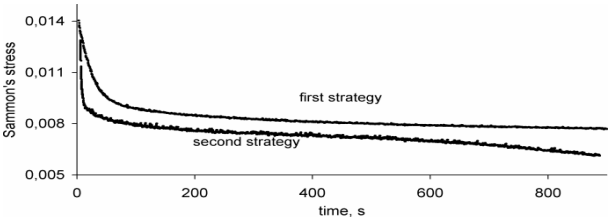

Fig 8. Dependence of the projection error on the computing time for the Iris dataset

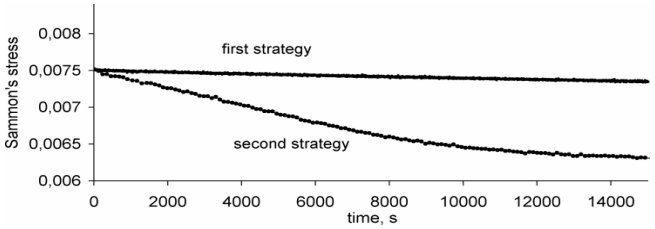

Fig 9. Dependence of the projection error on the computing time for randomly generated vectors

8. Analysis of the Medical Data Using the SAMANN Network

The results of the research are applied in solving some problems in practice. In this chapter, we investigate the human physiological data that describe the human functional state. A frequent problem in medicine is assignment of a health state to one of the known classes (for example, healthy or sick persons). Such an assignment is made by doctors as usual. It is of utmost importance to determine the bound of transit from one (normal) state to another (diseased), i.e., to determine the decision surfaces. We propose the way to find a certain area of parameter values, which correspond to the patients that should be thoroughly examined. The SAMANN method allows us to present this area in 2D form, which is preferable for the visual decision on the state of health of the patients.

A physiological data set has been analyzed. This data set consists of three groups: ischemic heart diseased patients (Group 1) (61 item), healthy persons (not going in for sports) (Group 2) (110 items), and sportsmen (Group 3) (161 item). The number of parameters is 18. At first the analyzed data were classified by the Naïve Baeys, classification tree, and support vector machine classifiers.

In the analysis of visualization of medical data, the SAMANN network was used. The SAMANN network was trained by two groups (ischemics and sportsmen), and the set of network weights was calculated. Now it is possible to decide where to place the third group data (healthy) in the final 2D configuration created by SAMANN. The third group shown to the network is mapped very fast and quite exactly without any additional calculations.

Fig 10(a) illustrates the projection of the data of Group 1 (ischemics) and Group 3 (sportsmen). Some groups are indicated: ischemics heart-diseased patients (assigned to ischemics by medics and by most classifiers); sportsmen, whom the majority of classifiers assigned to ischemics; sportsmen (assigned to sportsmen by medics and by most classifiers); ischemics, whom the majority of classifiers assigned to sportsmen. As seen in Fig 10(a), the majority of points, corresponding to ishemics, are on the one side of the picture, while the points, corresponding to sportsmen, are on the opposite side. However, these groups partly overlap. Therefore, it is reasonable to isolate the area of those intermix points of groups. One of the simplest ways is to connect the closest points, classified incorrectly or assigned t o the opposite class, by a broken line.

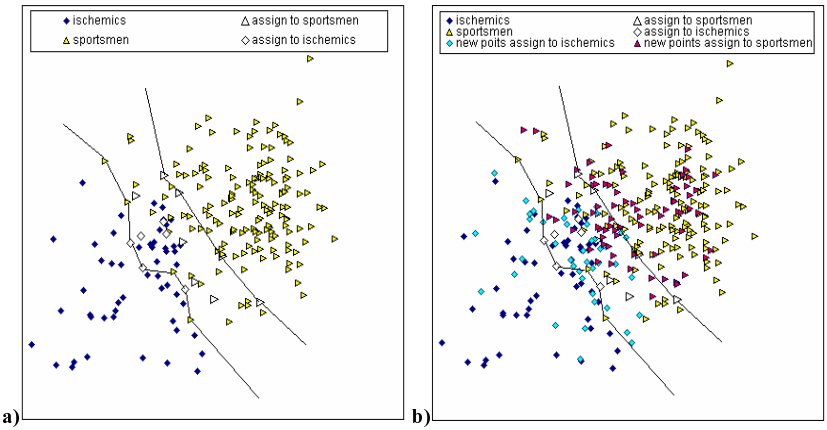

Fig 10. Projection of the physiological data: (a) groups 1 and 3; (b) all the 3 groups

It is difficult to tell by the visual image which class the points between the broken lines should be assigned to, and which of them correspond to a particular patient under investigation. These patients should be examined and observed more in detail. Fig 10(b) illustrates the points of all three groups. When visualizing the data, the points of Group 2 (healthy persons) are denoted with reference to the class they have been assigned to by most of the classifiers, i.e., it is indicated: healthy persons, assigned to Group 1 by the classifiers (circles); healthy persons, assigned to Group 3 by the classifiers (squares).

From the results of visual analysis we can draw the following conclusions:

1. The points, corresponding to persons not going in for sports, visually are arranged in the area of points, corresponding to the persons who go in for sports (on the right, in Fig 10(b)), are healthy and can go in for sports, because their measured parameters do not differ at all from the persons going in for sports.
2. The points, that correspond to persons not going in for sports visually are arranged in the area of points that correspond to ischemic patients (on the left, in Fig 10(b)), may possibly have some health problems and are not allowed to practice exercises without a thorough examination because their measured parameters do not differ at all from the ischemic patients.
3. It is also worthwhile observing and examining the sportsmen if the visual analysis shows that their points are arranged close to the points that correspond to ischemic patients, because there is a possibility for them to have some health disturbance.

9. General Conclusions

1. After a comprehensive analysis of the methods for the visualization of the multidimensional data has been made, a conclusion is drawn: using these methods, it is complicated to comprehend the data structure, it is almost impossible, if large data sets or data of large dimensions are analysed. It is easier to comprehend and to interpret the results, obtained by the projection methods, where multidimensional data are presented onto a lower dimension space. A more precise data structure is preserved using nonlinear projection methods. Nevertheless, the projection errors are inevitable.
2. It has been noticed that the SAMANN network that realizes Sammon's algorithm training depends on different parameters. The experiments, done in this work, show in what way the network training depends on the learning rate. The results of the experiments have shown that with an increase in the learning rate value, a better projection error is obtained.
3. The experimental investigation shows that the optimal value of the learning rate is in the interval (5;30). By selecting such values of the learning rate, a

significant economy of the computing time is possible for a fixed number of iterations (up to 3–5 times). Smaller values of the learning rate within the interval (0;1) guarantee a more stable convergence to the minimum of the mapping error. Some fluctuations are observed in the result when the rate is set to be larger. Meanwhile, these fluctuations are rather slight if the learning rate is in the interval (5;30).

4. Possibilities of using several computers for concurrent training of the network have been analyzed. When calculating by a parallel algorithm that splits the analyzed set into several parts, uses separate parts for independent network training and afterwards joins the weights obtained by averaging them after each training iteration (if all the possible pairs of vectors of a separate training set part are presented to the network once), we failed to improve the mapping results and (or) speed-up calculations. However, the results obtained allow us to draw conclusions that in creating new modifications of the parallel SAMANN algorithm, it is necessary to try to diminish data sending expenses and to rationally join the network weights computed by different processors.
5. A modification of the parallel algorithm has been proposed that divides the analyzed set into several parts, uses separate parts for independent network training, and, after a certain fixed number of iterations, selects the best projection results achieved regarding the whole data set, and completes network training with an aggregate data set. The research has shown that calculating by means of the parallel SAMANN algorithm modified in this way, it is possible to achieve better visualization results in a shorter time (as compared to the serial algorithm).
6. The results of the experiments have shown that it is possible to find a subset of the analyzed dataset such that in training the SAMANN network by which lower projection errors are obtained faster than by training with all the points of the set.
7. Using a parallel algorithm instead of a serial, the calculation expenses decrease due to the fact that when dividing the data among processors, the number of pairs of the vectors presented to the neural network decreases. This is due to the fact that the initial data set has to be split into separate parts.
8. Three strategies for retraining the neural network that visualizes multidimensional data have been proposed and investigated. The experiments have shown that it is expedient to take one vector from the primary dataset and the other from the new one at every step of training. This strategy yields smaller projection errors faster. The proposed strategies can also be applied in visualizing large-scale data sets. The experiments lead to the idea of minimizing the SAMANN neural network training time

by dividing the training process into two subprocesses: training of the network by a part of the data vectors, and then retraining of the network by the remaining part of the dataset. In this case, the training set will consist of some subsets. The smaller number of pairs of vectors will be used when training the network by the vectors of the subsets. This allows us to reach a similar visualization quality much faster.

9. The SAMANN algorithm has been applied in the analysis of medical data. The human physiological data that describe the human functional state have been analyzed. This data set consists of three groups: ischemic heart diseased patients, healthy persons (not going in for sports), and sportsmen. The results, obtained by the method, can be of use to medics for a preliminary diagnosis: healthy, unclear or sick persons.

List of Published Works on the Topic of the Dissertation

Articles in scientific publications from the ISI Web of Science list

1. MEDVEDEV, V.; DZEMYDA, G. Optimization of the SAMANN network training. *Journal of Global Optimization*, Springer, 2006, Vol. 35(4), p. 607–623. ISSN 0925-5001.
2. MEDVEDEV, V.; DZEMYDA, G. Speed Up of the SAMANN Neural Network Retraining. Artificial Intelligence and Soft Computing – ICAISC 2006, *Lecture Notes in Computer Science*, Springer, 2006, Vol. 4029, p. 94–103. ISSN 0302-9743.

Articles in the scientific publications from the ISI Proceedings Journal list

3. MEDVEDEV, V.; DZEMYDA, G. Retraining the Neural Network for Data Visualization. In *IFIP International Federation for Information Processing*, *Artificial Intelligence Applications and Innovations*, 2006, Vol. 204, p. 27–34. ISSN 1571-5736.
4. IVANIKOVAS, S.; MEDVEDEV, V.; DZEMYDA, G. Parallel Realizations of the SAMANN Algorithm. *Lecture Notes in Computer Science,* Adaptive and Natural Computing Algorithms, Springer, 2007, Vol. 4432, p. 179–188. ISSN 0302-9743.

Chapter in the reviewed book (IOS Press)

5. DZEMYDA, G.; KURASOVA, O.; MEDVEDEV, V. Dimension Reduction and Data Visualization Using Neural Networks. *Emerging Artificial Intelligence Applications in Computer Engineering. Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. Vol. 160, Frontiers in Artificial Intelligence and

Applications. Editors: I. Maglogiannis, K. Karpouzis, M. Wallace and J. Soldatos, IOS Press, 2007, p. 25–49. ISBN 978-1-58603-780-2.

Article in the reviewed book (Springer)

6. 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*, Springer, 2007, Vol. 4, p. 277–298. ISBN 978-0-387-36720-0.

In the other editions (in Lithuanian)

7. MEDVEDEV, V.; DZEMYDA, G. Vizualizavimo paklaidos lygiagrečiose SAMANN algoritmo realizacijose. *Lietuvos matematikos rinkinys*, 2004, T. 44, Spec. nr., p. 649–654. ISSN 0132-2818.
8. MEDVEDEV, V.; DZEMYDA, G. Daugiamačius duomenis vizualizuojančio neuroninio tinklo permokymo strategijos, *Informacijos mokslai*, 2005, T. 34, p. 263–267. ISSN 1392-0561.
9. MEDVEDEV, V.; DZEMYDA, G. Vizualizavimui skirto neuroninio tinklo mokymosi greičio optimizavimas. *Lietuvos matematikos rinkinys,* 2005, T. 45, Spec. nr., p. 426–431. ISSN 0132-2818.
10. MEDVEDEV, V.; DZEMYDA, G. Lygiagreti SAMANN vizualizavimo algoritmo realizacija. Iš *Informacinės technologijos 2004*, konferencijos pranešimo medžiaga. Kaunas: Technologija, 2004, p. 344–350. ISBN 9955-09-588-1.
11. MEDVEDEV, V.; DZEMYDA, G. SAMANN neuroninio tinklo mokymo problemos. Iš *Informacinės technologijos 2005*, konferencijos pranešimo medžiaga. Kaunas: Technologija, 2005, p. 394–399. ISBN 9955-09-588-1.

Short description about the author of the dissertation

1997–2001 – Studies at the Vilnius Pedagogical University, Faculty of Mathematics and Informatics – Bachelor of Mathematics.

2001–2003 – Studies at the Vilnius Pedagogical University, Faculty of Mathematics and Informatics – Master of Informatics.

2003–2007 – PhD studies at the Institute of Mathematics and Informatics, Systems Analysis Department.

e-mail: Viktor.M@ktl.mii.lt.

TIESIOGINIO SKLIDIMO NEURONINIŲ TINKLŲ TAIKYMO DAUGIAMAČIAMS DUOMENIMS VIZUALIZUOTI TYRIMAI

Tyrimų sritis ir problemos aktualumas. Šio darbo tyrimų sritis yra daugiamačių duomenų analizė, bei tų duomenų suvokimo gerinimo būdai. Duomenų suvokimas yra sudėtingas uždavinys, ypač kai duomenys nurodo sudėtingą objektą, kuris aprašytas daugeliu parametrų.

Pastaruoju metu stebima tendencija, kad MDS (daugiamačių skalių) tyrimus vykdantys mokslininkai dažnai atsiriboja nuo kitų metodų tyrimų, ar net ignoruoja tuos metodus. Antra vertus, kitų vizualizavimo metodų tyrimuose nėra lyginimų ar sąsajų su MDS tipo metodais. Šiame darbe siekiama praplėsti MDS tipo metodų realizacijas dirbtinių neuroninių tinklų taikymu, tuo būdu stiprinant ryšį tarp skirtingų vizualios duomenų analizės krypčių. Darbe nagrinėjami dirbtinių neuroninių tinklų algoritmai daugiamačiams duomenims vizualizuoti. Pasiūlyta specifinė mokymo taisyklė (SAMANN), kuri leidžia įprastam tiesioginio sklidimo neuroniniam tinklui realizuoti Sammono projekciją, mokymo be mokytojo būdu.

Analizuojamų daugiamačių duomenų projekcijos paklaidų minimizavimas naudojant dirbtinius neuroninius tinklus išlieka aktualia problema. Ta problema yra pagrindinė šioje disertacijoje sprendžiama **problema**.

Darbo tikslas ir uždaviniai. Pagrindinis disertacijos tikslas yra sukurti ir tobulinti metodus, kuriuos taikant būtų efektyviai minimizuojamos daugiamačių duomenų projekcijos paklaidos naudojantis dirbtiniais neuroniniais tinklais bei projekcijos algoritmais. Norint pasiekti šį tikslą, reikėjo išspręsti tokius uždavinius: 1) analitiškai apžvelgti daugiamačių duomenų vizualizavimo metodus; 2) ištirti dirbtinių neuroninių tinklų galimybes daugiamačiams duomenims vizualizuoti; 3) sukurti lygiagretųjį SAMANN vizualizavimo algoritmą; 4) pagreitinti (pagerinti) SAMANN tinklo mokymą ir permokymą; 5) surasti optimalias nagrinėjamo algoritmo mokymosi parametro reikšmes; 6) išanalizuoti naujų daugiamačių taškų vizualizavimo galimybes naudojantis dirbtiniais neuroniniais tinklais.

Tyrimų **metodikos** pagrindą sudaro naujų SAMANN neuroninio tinklo mokymo strategijų kūrimas ir jų eksperimentinis tyrimas.

Tyrimų objektas. Disertacijos tyrimų objektas yra dirbtiniai neuroniniai tinklai, skirti daugiamačių duomenų vizualizavimui. Su šiuo objektu yra betarpiškai susiję dalykai: 1) daugiamačių duomenų vizualizavimas; 2) dimensijos mažinimo algoritmai; 3) projekcijos paklaidos; 4) naujų taškų

atvaizdavimas; 5) vizualizavimui skirto neuroninio tinklo permokymo strategijos ir parametrų optimizavimas; 6) lygiagretieji skaičiavimai.

Mokslinis naujumas ir ginamieji teiginiai. Sukurtas lygiagretusis SAMANN algoritmas, realizuojantis daugiamačių duomenų vizualizavimą. Pasiūlytos ir ištirtos nagrinėjamo algoritmo permokymo strategijos. Eksperimentiškai nustatyta, kaip parinkti SAMANN tinklo mokymosi parametro reikšmę, kad algoritmas veiktų efektyviai.

Praktinė darbo reikšmė. Tyrimų rezultatai atskleidė naujas medicininių (fiziologinių) duomenų analizės galimybes. Tai leido sporto medicinos specialistams įvertinti nesportuojančiųjų sveikatos būklę ir jų galimybę sportuoti. Tyrimai atlikti pagal Lietuvos valstybinio mokslo ir studijų fondo prioritetinių Lietuvos mokslinių tyrimų ir eksperimentinės plėtros programą „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.

Darbo rezultatų aprobavimas ir publikavimas. Tyrimų rezultatai publikuoti 11 moksliniuose leidiniuose: 2 straipsniai leidiniuose, įtrauktuose į Mokslinės informacijos instituto pagrindinį (ISI Web of Science) sąrašą; 2 straipsniai leidiniuose, įtrauktuose į Mokslinės informacijos instituto konferencijos darbų (ISI Proceedings) sąrašą; 1 skyrius užsienio leidyklos (IOS Press) knygoje; 1 straipsnis užsienio leidyklos knygoje; 3 straipsniai Lietuvos periodiniuose leidiniuose; 2 straipsniai konferencijų pranešimų medžiagoje. Tyrimų rezultatai buvo pristatyti ir aptarti 9-iose nacionalinėse ir tarptautinėse konferencijose.

Darbo apimtis. Disertaciją sudaro devyni skyriai ir literatūros sąrašas. Bendra disertacijos apimtis 144 puslapiai, 86 paveikslai ir 1 lentelė.

Pirmame skyriuje išdėstytas disertacijos temos aktualumas, tyrimų sritis, suformuluotas tyrimo tikslas, pateikti tyrimo uždaviniai, aprašytas tyrimo objektas, darbo naujumas, praktinė vertė, darbo aprobavimas, pateiktas darbo publikacijų sąrašas, pristatyta darbo struktūra.

Antrame skyriuje pateikta tiesioginių daugiamačių duomenų vizualizavimo metodų ir projekcijos metodų analizė. Tiesioginiuose vizualizavimo metoduose kiekvienas daugiamačio vektoriaus parametras pateikiamas tam tikra vizualia forma. Projekcijos metodų tikslas – pateikti daugiamačius duomenis mažesnės dimensijos erdvėje taip, kad būtų kiek galima tiksliau išlaikyta tam tikra duomenų struktūra, ir palengvinti didelės dimensijos duomenų interpretavimą bei apdorojimą.

Trečiame skyriuje pateiktos esminės dirbtinių neuroninių tinklų koncepcijos, kurios yra reikalingos analizuojant neuroninių tinklų galimybes vizualizuoti daugiamačius duomenis: dirbtinio neurono modelis, vienasluoksnio perceptrono sandara, daugiasluoksnių tiesioginio sklidimo neuroninių tinklų sandara, radialinių bazinių funkcijų neuroniniai tinklai, saviorganizuojantys neuroniniai tinklai.

Ketvirtame skyriuje analizuotos dirbtinių neuroninių tinklų galimybės vizualizuoti daugiamačius duomenis, kadangi klasikiniai vizualizavimo metodai kartais yra nepajėgūs susidoroti su savo užduotimis.

Penktame skyriuje analizuota specifinė „klaidos sklidimo atgal“ mokymo taisyklė, SAMANN, kuri leidžia įprastam tiesioginio sklidimo neuroniniam tinklui realizuoti Sammono projekciją mokymo be mokytojo būdu. Buvo nagrinėjama SAMANN tinklo mokymo proceso priklausomybė nuo mokymosi parametro reikšmės. Nustatyta optimali mokymosi parametro reikšmė.

Šeštame skyriuje pasiūlytos lygiagrečios SAMANN algoritmo realizacijos, leidžiančios tinklo mokymui vienu metu naudoti keletą procesorių. Pateikiami gauti rezultatai ir išvados.

Septintame skyriuje analizuojamos SAMANN tinklo galimybės vizualizuoti naujus duomenis. Pasiūlytos daugiamačius duomenis vizualizuojančio neuroninio tinklo permokymo strategijos ir atlikta jų analizė.

Aštuntame skyriuje SAMANN tinklas taikytas medicininių (fiziologinių) duomenų analizei.

Devintame skyriuje pateiktos disertacijos išvados.

Bendrosios išvados

1. Tiesioginio vizualizavimo metodų analizė parodė, kad naudojant šiuos metodus suprasti duomenų struktūrą yra gana sudėtinga, ypač esant didesnei duomenų dimensijai arba analizuojant didelės apimties duomenų aibę. Daug lengviau suvokti ir interpretuoti rezultatus, gautus projekcijos metodais, kuriuose daugiamatis vektorius transformuojamas į mažesnės dimensijos vektorių. Lyginant tiesinės ir netiesinės projekcijos metodus, tikslesnė duomenų struktūra išlaikoma naudojant netiesinės projekcijos metodus. Bet ir čia duomenų vizualizavimo iškraipymai yra neišvengiami.
2. Analizuojant mokymo be mokytojo „klaidos sklidimo atgal“ SAMANN neuroninį tinklą, nustatyta, kad projekcijos paklaida ir tinklo konvergavimas priklauso nuo pasirinktų parametrų reikšmių. Eksperimentai parodė, kad, kuo didesnė mokymosi parametro reikšmė, tuo greičiau pavyksta pasiekti gerus vizualizavimo rezultatus. Tačiau, didėjant mokymosi parametro reikšmei, didėja paklaidos svyravimai.
3. Tiriant kelias duomenų aibes, nustatyta, kad optimali SAMANN tinklo mokymosi parametro reikšmė yra intervale (5;30). Pasirenkant tokias

mokymosi parametro reikšmes, galima žymiai sumažinti skaičiavimų trukmę (iki 3−5 kartų ir net daugiau) ir gauti gerus vizualizavimo rezultatus per trumpesnį laiką, esant fiksuotam iteracijų skaičiui. Mažos mokymosi parametro reikšmės intervale (0;1) garantuoja stabilų projekcijos paklaidos mažėjimą didėjant iteracijų skaičiui. Tuo tarpu, kai mokymosi parametro reikšmė pasirenkama didesnė, pastebimi tam tikri paklaidos svyravimai. Tačiau šie svyravimai yra pakankamai maži, kai mokymosi parametro reikšmė pasirenkama iš intervalo (5;30).

4. Ištirtos galimybės tinklo mokymui vienu metu naudoti keletą kompiuterių. Skaičiuojant lygiagrečiuoju SAMANN algoritmu, kuris analizuojamą aibę dalina į kelias dalis, vykdo nepriklausomai tinklo mokymą atskiromis dalimis, o po to gautus svorius apjungia juos vidurkinant po kiekvienos mokymo iteracijos, nepavyko pagerinti atvaizdavimo rezultatų ir (arba) pagreitinti skaičiavimų. Tačiau, gauti rezultatai, leidžia daryti išvadas, kad kuriant naujas lygiagrečias SAMANN algoritmo modifikacijas būtina siekti mažinti duomenų persiuntimo kaštus ir racionaliai atlikti atskirais procesoriais apskaičiuotų tinklo svorių apjungimą.
5. Pasiūlyta lygiagrečiojo algoritmo modifikacija, kuri analizuojamą aibę dalina į kelias dalis, vykdo nepriklausomai tinklo mokymą atskiromis dalimis, o po tam tikro fiksuoto iteracijų skaičiaus pasirenka geriausią pasiektą projekcijos rezultatą visos duomenų aibės požiūriu ir baigia mokyti tinklą pilna duomenų aibe. Tyrimai parodė, kad skaičiuojant lygiagrečiuoju SAMANN algoritmu galima pasiekti geresnius vizualizavimo rezultatus per trumpesnį laiką (lyginant su nuosekliuoju algoritmu).
6. Eksperimentai parodė, kad galima rasti tokį analizuojamos duomenų aibės poaibį, kuriuo mokant SAMANN tinklą, mažesnės projekcijos paklaidos gaunamos greičiau, negu tinklo mokymui naudojant visus aibės taškus.
7. Pereinant nuo nuoseklioji algoritmo prie lygiagretaus algoritmo, skaičiuojamieji kaštai sumažėja dėl to, kad skirstant duomenis tarp procesorių, sumažėja vektorių porų, pateikiamų neuroniniam tinklui, skaičius.
8. Pasiūlytos ir ištirtos trys daugiamačius duomenis vizualizuojančio neuroninio tinklo permokymo strategijos. Eksperimentai parodė, kad iš pasiūlytų strategijų geriausia yra tokia strategija, kai kiekvieno mokymo žingsnio metu vienas vektorius imamas iš senos duomenų aibės, o kitas iš naujos. Ši strategija duoda mažesnes projekcijos paklaidas lyginant su kitomis strategijomis. Ji leidžia sumažinti skaičiavimų trukmę tam pačiam rezultatui pasiekti. Visos trys pasiūlytos permokymo strategijos gali būti taikomos didelės apimties duomenų aibių vizualizavimui. Galimas būdas minimizuoti SAMANN tinklo mokymo laiką yra mokymo proceso

padalinimas į du subprocesus: tinklo mokymas analizuojamos duomenų aibės dalimi, vėliau – tinklo permokymas likusia duomenų aibės dalimi arba visa aibe. Šiuo atveju tinklo mokymo proceso pagrindinę dalį užimtų mokymas nepilna duomenų aibe, o tai iš esmės taupytų skaičiavimo laiką.

9. SAMANN algoritmas buvo pritaikytas medicininių (fiziologinių) duomenų analizei. Fiziologinių duomenų aibė sudaryta iš trijų grupių vyrų širdies funkcinių rodiklių rinkinių: vyrai, sergantys išemine širdies liga, ne sportininkai (vyrai, kuriems liga nebuvo diagnozuota) ir profesionalūs sportininkai. Analizė leido sporto medicinos specialistams įvertinti nesportuojančių vyrų sveikatos būklę ir jų galimybę sportuoti. SAMANN algoritmo privalumas vizualizuojant tirtus medicininius duomenis yra tai, kad nauji taškai (duomenys apie nesportuojančius vyrus) „randa“ savo vietas tarp jau atvaizduotų duomenų su žinomomis charakteristikomis.

Viktor Medvedev

RESEARCH OF MULTIDIMENSIONAL DATA VISUALIZATION USING FEED-FORWARD NEURAL NETWORKS

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

Viktor Medvedev

TIESIOGINIO SKLIDIMO NEURONINIŲ TINKLŲ TAIKYMO DAUGIAMAČIAMS DUOMENIMS VIZUALIZUOTI TYRIMAI

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

2007 12 07. 1,5 sp. l. Tiražas 100 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11
LT-10223 Vilnius, *http://leidykla.vgtu.lt*
Spausdino UAB „Baltijos kopija“,
Kareivių g. 13B, 09109 Vilnius, *www.kopija.lt*