Jolita BERNATAVIČIENĖ

METHODOLOGY OF VISUAL KNOWLEDGE DISCOVERY AND ITS INVESTIGATION

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

1494-M

VGTU LEIDYKLA TECHNIKA

Vilnius 2008

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Jolita BERNATAVIČIENĖ

METHODOLOGY OF VISUAL KNOWLEDGE DISCOVERY AND ITS INVESTIGATION

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

VGTU
Vilnius LEIDYKLA TECHNIKA 2008

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2004–2008.

Scientific Supervisor

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

Consultant

Prof Dr Habil Vydūnas ŠALTENIS (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:

Assoc Prof Dr Vitalijus DENISOVAS (Klaipėda University, Physical Sciences, Informatics – 09P),

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

Prof Dr Habil Alvydas PAUNKSNIS (Kaunas University of Medicine, Biomedical Sciences, Medicine – 07B),

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

Opponents:

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

Prof Dr Habil Rimvydas SIMUTIS (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 Engineering in the Conference and Seminars Center of the Institute of Mathematics and Informatics at 2 p. m. on 20 June 2008.

Address: Goštauto g. 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 20 May 2008.

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

Jolita BERNATAVIČIENĖ

VIZUALIOS ŽINIŲ GAVYBOS METODOLOGIJA IR JOS TYRIMAS

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

Vilnius LEIDYKLA TECHNIKA 2008

Disertacija rengta 2004–2008 metais Matematikos ir informatikos institute.

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

Konsultantas
prof. habil. dr. Vydūnas ŠALTENIS (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:
doc. dr. Vitalijus DENISOVAS (Klaipėdos universitetas, fiziniai mokslai, informatika – 09P),
doc. dr. Regina KULVIETIENĖ (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),
prof. habil. dr. Alvydas PAUNKSNIS (Kauno medicinos universitetas, biomedicinos mokslai, medicina – 07B),
prof. habil. dr. Antanas ŽILINSKAS (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).

Oponentai:
doc. dr. Antanas Leonas LIPEIKA (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T),
prof. habil. dr. Rimvydas SIMUTIS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija bus ginama viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2008 m. birželio mėn. 20 d. 14 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 2008 m. gegužės 20 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“ 1494-M mokslo literatūros knyga.

General characteristic of the dissertation

Topicality of the problem. The research area of this work is the process of knowledge discovery from multidimensional data and the ways of improving data recognition and perception.

We constantly face multidimensional data in technics, medicine, economics, ecology, and many other areas. With the progress in technologies, perception improvement of computers and software, volumes of accumulated data are rapidly increasing. However, there still remains a wide gap between data collection and storage and their perception as well application of the acquired knowledge in solving practical problems. Perception of multidimensional data is the result of a long and complicated knowledge discovery process. This process is transition from a large analysed data set to specific data from which information is extracted and knowledge about the structure, new connections of the investigated data and the groups are formed, which will influence further decision making.

Several stages of the knowledge discovery process are considered in detail in literature, however there is no single integral methodology that includes all the stages of knowledge discovery. It can enable a researcher to join this information with an expert's experience and to establish a knowledge bank which will be helpful to solve the problems posed above.

The problem under consideration is assurance of integrity of the knowledge discovery process.

Aim and tasks of the work. The key aim of the dissertation is to develop and explore the methodology of knowledge discovery by visual methods that would allow us to improve the efficiency of data analysis. With a view to achieve this aim, we had to solve the following problems: (1) to analytically overview the methods of data mining and knowledge discovery; (2) to investigate the process of knowledge discovery; to look over and compare the existing models of this process; to examine the possibilities of applying visualization in the knowledge discovery process; to propose and study a model for a multidimensional data visualization on which the developed methodology is based; (3) to explore the selected algorithms, used in the visual knowledge discovery process, and to create more efficient their modifications; to consider the visualization possibilities of the newly obtained multidimensional data and to improve the efficiency of the methods employed; (4) to propose and investigate the ways of changing the data arrangement geometry with the view of a more accurate data projection on the plane; (5) to apply the developed methodology in the analysis of medical and physiological data.

Research object. The research object of the dissertation is the process of visual knowledge discovery from multidimensional data. The following subjects are directly connected with this object: (1) formation of a primary set of multidimensional data; (2) algorithms for clusterization, visualization, and classification; (3) evaluation of the results obtained by data mining methods; (4) mapping of the new multidimensional data; (5) decision making and generalization of the knowledge obtained regarding the analysis results.

Scientific novelty. The methodology for knowledge discovery by visual methods has been developed that enables us to make an exhaustive and informative analysis of the data under investigation. The ways of improving the efficiency of the relative multidimensional scaling (MDS) algorithm have been proposed: (1) strategies for selecting basic vectors have been created; (2) initialization problems in the relative MDS algorithm have been explored, and the best way of initializing two-dimensional (2D) vectors established; (3) the way of selecting the optimal number of basic vectors has been proposed. The transformation of distances between multidimensional data has been developed that increases the visualization quality: it better exposes data clusters and less distorts the structure of multidimensional data; (4) an approach of preliminary evaluation of the health state on the basis of physiological data analysis has been proposed, using this methodology.

Analytical analysis, generalization, and experimental study make up the basis of the research **methodology.**

Defended propositions

1. Systematization of the knowledge discovery process allows all-round evaluation and application of the abilities of visualization methods and means to increase the efficiency of data analysis.
2. It is possible to improve the efficiency of relative multidimensional scaling by selecting the proper number of basic vectors, the strategy for selection of basic vectors as well as the way of initializing two-dimensional vectors.
3. There is a possibility to raise the visualization quality of multidimensional vectors, applying the transformation of adjustment of multidimensional data distances.
4. The developed methodology for visual knowledge discovery can be applied in a preliminary diagnosis of the health state.

Practical value. The research results displayed new capabilities of medical and physiological data analysis. They enabled sports medicine experts to evaluate the health state of those not going in for sports and their ability to go in for sports. The investigations have been pursued in line with:

- 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.
- the Lithuanian State Science and Studies Foundation project “Information technology tools of clinical decision support and citizens wellness for e.Health system, Info Health (No. B-07019)”.

Approbation and publications of the research. The main results of this dissertation were published in 9 scientific papers: 1 article in a journal abstracted in Thomson ISI Web of Science database; 2 articles in scientific publications indexed in Thomson ISI Proceedings database; 3 articles in journals indexed in international databases approved by Science Council of Lithuania; 3 articles in the proceedings of scientific conferences. The main results of the work have been presented and discussed in 5 international and 4 national conferences.

The scope of the scientific work. The work is written in Lithuanian. It consists of 5 chapters, and the list of references. There are 116 pages of the text, 44 figures, 12 tables and 156 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. The role of visual analysis in knowledge discovery

This chapter presents the analytic investigation of the data discovery problems solved as well as methods for solving these problems. Systematized and considered visualization methods, based on different ideas, are universal or oriented to specific data. The data mining methods that can be combined with visualization methods and used in the process of knowledge discovery are surveyed as well. The knowledge discovery process is also discussed as well as four models of this process are overviewed and compared. We have shown that their principal difference lies in the level of specification. After generalizing we obtain a six-step scheme that includes all the stages of knowledge discovery and a proper place of visualization is set in each stage of knowledge discovery.

The analysis has shown that visualization plays an important role in the process of knowledge discovery, however its usage is rather fragmentary. One needs here an integrated view point that embraces all the stages the knowledge discovery process.

3. Extension of the abilities of visual knowledge discovery

We deal here with methodology of visual knowledge discovery, the ways of improving the efficiency of the relative multidimensional scaling, and the algorithm for adjusting mutual distances of multidimensional data, which is used in the visualization of multidimensional data.

3.1. Methodology of visual knowledge discovery

Having evaluated all the stages of knowledge discovery process, we have established that a fragmentary use of visualization does not exhaust all the abilities that visualization can ensure. It should be integrated into all the stages of the knowledge discovery process.

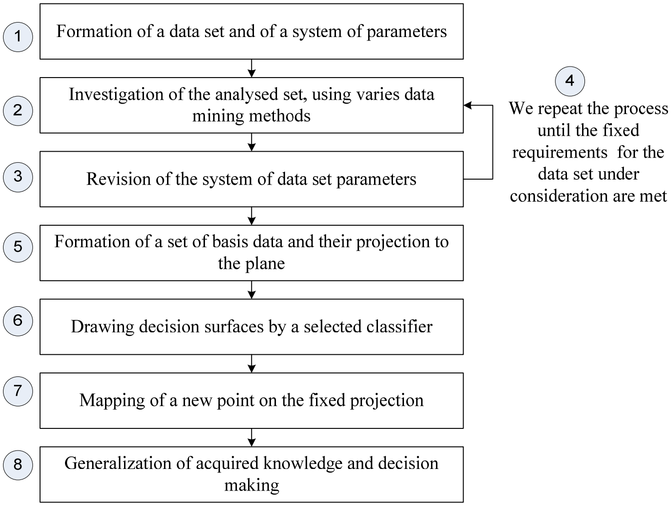

1
Formation of a data set and of a system of parameters
2
Investigation of the analysed set, using varies data mining methods
3
Revision of the system of data set parameters
4
We repeat the process until the fixed requirements for the data set under consideration are met
5
Formation of a set of basis data and their projection to the plane
6
Drawing decision surfaces by a selected classifier
7
Mapping of a new point on the fixed projection
8
Generalization of acquired knowledge and decision making

Fig 1. The scheme of the visual knowledge discovery process

The methodology for visual knowledge discovery has been proposed in this chapter on the basis of which an exhaustive visual data analysis is made. In the proposed methodology, graphically illustrated in Fig 1, visualization is integrated into the stages of the knowledge discovery process.

3.2. Improvement of the efficiency of the relative multidimensional scaling

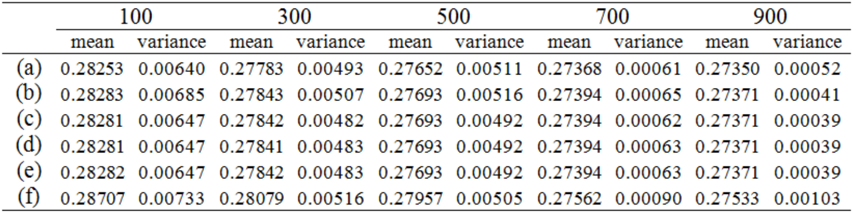
Multidimensional scaling (MDS) is a group of methods that project multidimensional data to a low (usually two) dimensional space and preserve the interpoint distances among data as much as possible. Let us have vectors $X^i = (x_1^i, x_2^i, ..., x_n^i)$, $i = 1, ..., m$ ($X^i \in R^n$). The pending problem is to get the projection of these *n*-dimensional vectors X^i, $i = 1, ..., m,$ onto the plane R^2. Two-dimensional vectors $Y^1, Y^2, ..., Y^m \in R^2$ correspond to them. Here $Y^i = (y_1^i, y_2^i)$, $i = 1, ..., m.$ Denote the distance between the vectors X^i and X^j by d_{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 (1) is minimized $E_{MDS} = \sum_{i<j}^{m} (d_{ij}^* - d_{ij})^2$. The SMACOF algorithm for the minimization of the stress function based on iterative majorization has been used. It is one of the best optimization algorithms for this type of minimization problem. This method is simple and powerful, because it guarantees a monotone convergence of the stress function.

Relative MDS. In classification tasks, it may be interesting to see where a new data point "falls" among the known cases and to discover the class topology of its neighbouring known cases to get an insight on how a classifier would classify this new point. The MDS is a topology preserving mapping, but it does not offer an opportunity to project new points on the existing set of mapped points. To get a mapping that presents the previously mapped points together with the new ones requires a complete re-run of the MDS algorithm on the new and the old data points. Let us denote the number of known data points by N_{fixed}, the number of new data points by N_{new}, the total number of points considered during the mapping by N_{total} ($N_{total} = N_{fixed} + N_{new}$), the set of known data points by F (it will be called a basis vector set), the set of new data points by M. The algorithm scheme is as follows: (1) map set F using the MDS mapping (the number of fixed points is equal to N_{fixed}); (2) map set

M with respect to the mapped set F using the relative MDS mapping (the number of new points is equal to N_{new}). The relative MDS mapping differs from the normal MDS by the fact that during the minimization of the stress function only the points from set M are allowed to move, while the points from set F are kept fixed. This is achieved by modifying the stress function so that it sums only over the distances that change during iterations, i.e., the distances between the fixed and moving points, and interpoint distances between the moving points. The stress function is rewritten as $E_{Relative\_MDS} = \sum_{i<j}^{N_{new}} (d_{ij}^{*} - d_{ij})^2 + \sum_{i=1}^{N_{new}} \sum_{j=N_{new}+1}^{N_{total}} (d_{ij}^{*} - d_{ij})^2$. In experiments, we use the Quasi-Newton algorithm to minimize $E_{Relative\_MDS}$.

It is necessary to find out which way of initialization to choose in order to project the remaining vectors on the fixed 2D map of basis vectors using the relative mapping. We have chosen 6 different initialization ways: (a) the matrix *A*[1x*n*] of average and rotation matrix *T*[*n*x2], obtained by using PCA in basis vector initialization, are saved; 2D coordinates of the remaining vectors are initialized by the formula: $Y_i = (X_i - A)T$, $i = 1,...,m$; (b) the initial coordinates of the vector from the remaining vector set are chosen as a 2D projection of the closest basis vector; (c), (d), (e) a random 2D vector, generated in the area of projection of the nearest basis vector, is attributed to the initial coordinates of a vector from the remaining vector set ((c) radius of the area $r = 0.01$; (d) $r = 0.1$; (e) $r = 1$); (f) a random 2D vector, generated in the area covered by all the 2D projections of basis vectors, is attributed to the initial coordinates of a vector from the remaining vector set.

Table 1. Experimental results for the *Gaussian* [2729, 10] data set



	100		300		500		700		900	
	mean	variance	mean	variance	mean	variance	mean	variance	mean	variance
(a)	0.28253	0.00640	0.27783	0.00493	0.27652	0.00511	0.27368	0.00061	0.27350	0.00052
(b)	0.28283	0.00685	0.27843	0.00507	0.27693	0.00516	0.27394	0.00065	0.27371	0.00041
(c)	0.28281	0.00647	0.27842	0.00482	0.27693	0.00492	0.27394	0.00062	0.27371	0.00039
(d)	0.28281	0.00647	0.27841	0.00483	0.27693	0.00492	0.27394	0.00063	0.27371	0.00039
(e)	0.28282	0.00647	0.27842	0.00483	0.27693	0.00492	0.27394	0.00063	0.27371	0.00039
(f)	0.28707	0.00733	0.28079	0.00516	0.27957	0.00505	0.27562	0.00090	0.27533	0.00103

With view to obtain a more precise projection of the whole data set, we suggest applying the PCA algorithm in the initialization of 2D vectors, corresponding to the remaining *n*-dimensional points. However, the differences of visualization results, obtained by all the five investigated ways of initialization, are not so significant. The worst way of initialization is a generation of random 2D vectors in the area covered by all the 2D projections of basis vectors. Table 1 illustrates these results.

In visualizing data sets the dimensionality of which is larger than 5 and that contain more than 3000 vectors, it is more reasonable to use the relative MDS algorithm. Under the above mentioned conditions, the relative MDS algorithm gives precise mapping and saves much computing time as compared with the standard MDS algorithm. Therefore, in the case of limited computing time, the projection by the relative MDS algorithm will be better than that by the standard MDS algorithm. The experiments are done with the following number of the data set vectors chosen randomly: $N_{fixed} = 100, 200, ..., 1500.$ Each experiment has been repeated for 10 times with a different set of basis vectors, the projection error and calculating time being estimated. These errors and time, obtained in 10 experiments, are averaged and presented in Fig 2 (Relative MDS, grey line). The projection error and calculation time in each iteration are presented in Fig 2 (standard MDS, black line) for each data set.

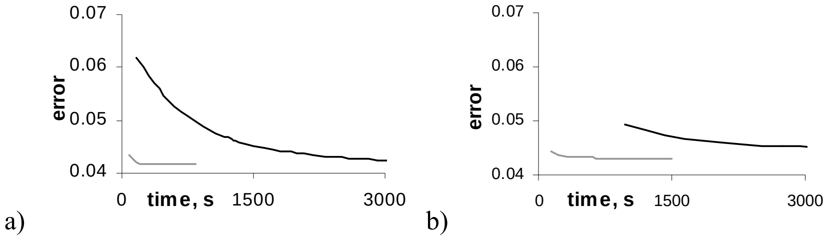


Fig 2. Dependence of the projection error on the computing time:
(a) *ellipsoidal* [3140, 50] data set; (b) *abalone* [4177, 7] data set

The larger dimensionality of visualized vectors requires the larger number of the basis vectors. The dependence of the projection error on the number of the basis vectors N_{fixed} is presented in Fig 3. It shows that the averaged projection error E constantly decreases, when N_{fixed} increases. The averaged projection error E stabilizes itself at $N_{fixed} \approx 700$ for small data sets (from 1000 to 3000 vectors) and at $N_{fixed} \approx 900$ for large data sets (more than 3000) (Fig 3). By increasing the number N_{fixed} even more the projection error changes insignificantly. With an increase in number of basis vectors, the mean value of E decreases; the variance of the projection error decreases significantly.

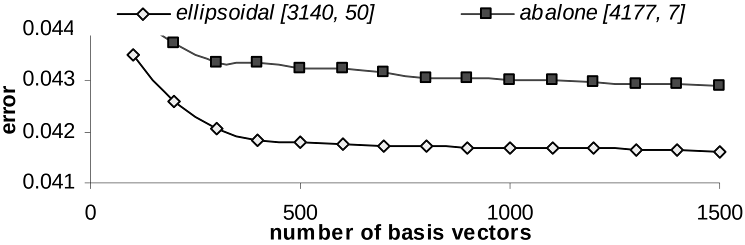


Fig 3. Dependence of the projection error on the number of the basis vectors

The basis vectors should be selected so that the basis vectors were distributed as uniformly as possible all over the data set, which shows better results of obtained visualization.

3.3. Correction of distances in the visualization of multidimensional data

The method of correction of interpoint distances is presented. The basic idea is to change the distances according to the distribution of distances in high dimensional areas.

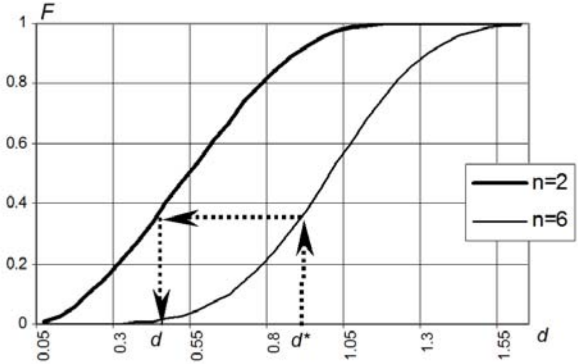


Fig 4. Example of the correction coefficient evaluation. The distribution functions of distances for 2D and 6D spaces are presented; $d^*[n=6]$ is the value of distance for a 6D space, $d[n=2]$ is the corresponding distance after correction for a 2D space

The correction must be such that the distances $d^*[n]$ between the uniformly distributed points in an n-dimensional hypercube would have the same distribution of distances $d[n=2]$ as in a 2D space (Fig 4). To this end, the distances $d^{*}[n]$ are shortened multiplying them by the corresponding correction coefficient $k_n = \frac{d}{d^{*}}$. Here $F_n(d^{*}) = F_2(d)$, $F_n(d^{*})$, $F_2(d)$ are distribution functions for n- and 2D spaces.

Thus we obtain the correction coefficient values for various distances d_n in an n-dimensional space (the distances are normalized in the interval [0,1]).

The function $k_n = 1 - \exp(-c_1(d^{*} - c_2))$ approximates these dependences well enough. The values of coefficients c_1 and c_2 are presented in Table 2. The values of the correction coefficients were calculated for various data dimensionalities. In order to demonstrate the quality of visualization, the experiments have been done on some datasets.

Table 2. Values of the coefficients c_1 and c_2

Dimension n	Value of the coefficient c_1	Value of the coefficient c_2
4	1.4	0.04
5	1.18	0.16
6	1.05	0.25

The proposed corrections are simple enough. It is necessary to multiply the pair wise distances in a multidimensional space by the respective corresponding correction coefficients that are different for various dimensionalities. The table of the correction coefficient values and approximated functions is presented for several dimensionalities.

4. Visual knowledge discovery by analysing physiological data

The human physiological features are often measured at certain time moments and some time series are obtained. A physiological data set has been analyzed. This data set consists of two groups: ischemic heart diseased patients (61 item), and sportsmen (161 item). The parameters to be used to describe the health state are derived from these data series. In this work, we discuss two ways of a generation of data series: the use fractal dimensions and polynomial approximation. Our investigation consists of four parts:

(1) *Comparative analysis of two parameter systems integrating some data mining methods – classification and visualization*. The research, based on the physiological data set of the sportsmen and ischemic heart diseased men, has

shown that the polynomial approximation is better than the fractal dimensions. The classification results, obtained by Naïve Bayes (NB), k nearest neighbour (kNN), Classification tree (CT), Support vector machine (SVM) classifiers, are presented in Table 3. Projections of the vectors, consisting of the parameters of: (Fig 5a) the capacity dimension (*n*=4); (Fig 5b) the polynomial parameter system (*n*=17), are presented. Projections are obtained using MDS SMACOF algorithm. As we see in Fig 5a, the data of both groups are overlapping enough. The misclassified points (SVM classifier is used), corresponding to the sportsmen (unfilled squares), "escape" from their group, i.e., they are farther from the majority of the points, corresponding to the sportsmen; they mix up among the points, corresponding to the ischemics. Doctors have some doubt whether these sportsmen really possess no symptoms of the ischemic heart-disease. As we see in Fig 5b, the visual overlapping of the two groups is lesser as compared with Fig 5a.

Table 3. Classification statistics when the vectors consisting of the parameters of polynomial aproximation and capacity fractal dimensions are classified

classi-fier	**parameter system of polynomial aproximation**			**parameter system of capacity fractal dimension**		
	CGA	**Spec.**	**Sens.**	**CGA**	**Spec.**	**Sens.**
NB	0.9152	0.9441	**0.8413**	0.8265	0.8820	0.6825
kNN	0.8840	0.9503	0.7143	0.8968	0.9317	**0.8095**
CT	0.8887	0.9503	0.7302	0.8573	0.9068	0.7302
SVM	0.9241	0.9627	**0.8254**	0.8970	0.9379	**0.7937**

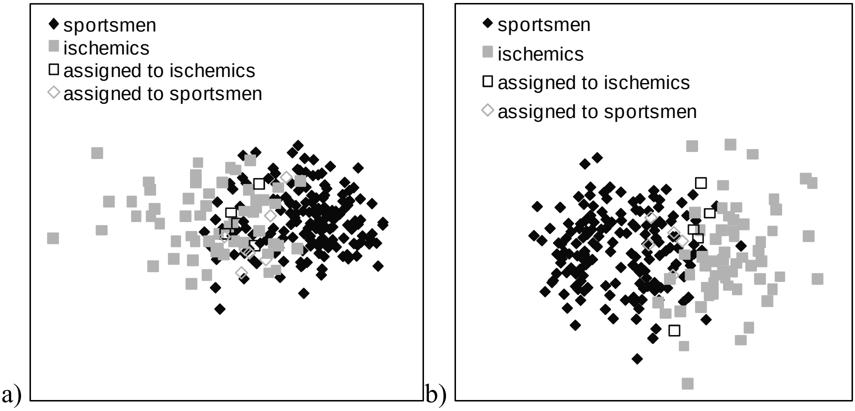


Fig 5. Projections of the vectors consisting of the parameters of: (a) the capacity dimension (*n*=4); (b) the new parameter system (*n*=17)

(2) *analysis of parameter estimation based on the correlation matrix analysis and visual presentation;* This analysis of the polynomial approximation-based parameter system allowed us to refuse heart rate parameters without loss of classification efficiency.

(3) *determination of the decision surface of some classes;* when classifying multidimensional data, it is important to determine a decision surface that separates the data into two classes. A preliminary diagnosis of a new patient can be determined according to the location of the point, corresponding to this patient, with respect to the decision surface. It is impossible to comprehend the decision surface in the *n*-dimensional space (when $n > 3$). So we suggest making a decision according to the data presented visually. We eliminate some data from the training data set, and interpret them as new points. First it is necessary to compare the quality of classification of the multidimensional data $X^1, X^2, ..., X^m \in R^{17}$ (first line of Table 4), and their 2D projection $Y^1, Y^2, ..., Y^m \in R^2$ (second line of Table 4), obtained by the MDS algorithm. The SVM is used for experiments.

Table 4. The classification quality

space	General accuracy	Specificity	Sensitivity
R^{17}	0.925	0.967	0.817
R^2	0.902	0.954	0.767

While comparing the results of the classification of 17-dimensional data with that of 2D data, we can state that the mapping of the multidimensional data onto the plane but slightly distorts the data structure and the changes in the classification quality are inessential. We use "SVM Toolbox for Matlab" to determine decision surfaces by the SVM. In Fig 6, the projections of the multidimensional data, support vectors and decision surfaces are presented: the points, corresponding to ischemics, are marked by filled squares; the points, corresponding to sportsmen, are marked by filled rhombi; the support vectors are marked by unfilled squares or rhombi (total 55); the bold line marks the decision surface, the thin solid line denotes the decision boundary of sportsmen, and the dashed line marks the decision boundary of ischemics.

(4) *mapping the data of the new patients among data points,* corresponding to patients with a known diagnosis. Mapping the data of the new patients among data points, corresponding to the patients with a known diagnosis, is very helpful to medical doctors in decision making on a patient's health state

and his capability to go in for sports and to discover the troubles in sportsman's heart action at the initial stage.

For example, in Fig 6, the points, marked by circled crosses, correspond to patients whose diagnosis has not yet been made. We can make a preliminary diagnosis of these patients as follows: one patient (No. 1) is healthy and can go in for sports, because the point, corresponding to this patient, falls to "the sportsmen area" (to the left of the bold and thin solid lines); ischemic heart disease is suspected in one patient (No. 2), because the point, corresponding to this patient, falls to "the ischemics area" (to the right of the dashed line) and this patient should be thoroughly examined immediately. The problems in the heart action may be suspected in a patient (No. 3), because the point, corresponding to this patient, falls to the area to the right of the bold line, but it is to the left from the dashed line, so this patient should be thoroughly examined to rate the seriousness of the health state.

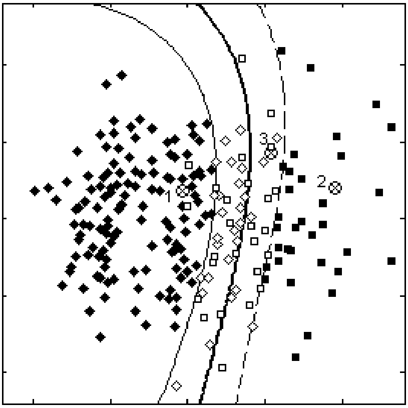


Fig 6. The new points of unknown classes, mapped onto the fixed projection

It is namely the application of the polynomial approximation-based system that allowed us to discover two subclusters among the sportsmen. The hypothesis is that one subcluster consists of the really healthy sportsmen, and a special attention should be paid to the cases of the second subcluster in terms of risk of the ischemic heart disease. Moreover, using this scheme, we can evaluate the possible extent of disease. Depending on the location of the points, corresponding to new patients, doctors can decide whether the new patients may be assigned to one of the known classes or the decision remains questionable.

5. General conclusions

1. Systematization of visual knowledge discovery process allowed an all-round evaluation and application of abilities of visualization methods and measures with a view to increase the efficiency of data analysis.
2. After a detailed exploration of the relative MDS algorithm, we can draw the following conclusions:
   - Visualization results depend on the selection strategy of basic vectors (the more accurate distribution of the basic vectors on the entire set, the more accurate projection is obtained), on the number of basic vectors, and on the way of initialization of two dimensional vectors.
   - With a view to obtain a more precise projection of the whole data set, we suggest applying the PCA algorithm in the initialization of two-dimensional vectors, corresponding to the remaining *n*-dimensional points. However, the differences of visualization results obtained by all the five investigated ways of initialization are not so significant. The worst way of initialization is a generation of random two-dimensional vectors in the area covered by all the two-dimensional projections of basis vectors.
   - In visualizing data sets the dimensionality of which is larger than 5 and that contain more than 3000 vectors, it is more reasonable to use the relative MDS algorithm. Under the above mentioned conditions, the relative MDS algorithm gives precise mapping and saves much computing time as compared with the standard MDS algorithm. Therefore, in the case of limited computing time, the projection by the relative MDS algorithm will be better than that by the standard MDS algorithm.
   - The larger dimensionality of visualized vectors requires the larger number of the basis vectors. When the number of the basis vectors increases, a more precise projection is obtained. However, too large number of the visualized basis vectors extends the computing time. The optimal number of basis vectors ranges from 700 to 1000 for small data sets (up to 3000), while for larger than 3000 data sets it ranges from 900 to 1500. With an increase of number of basis vectors, the mean value of the projection error decreases; the variance of the projection error decreases significantly.
3. Transformation of mutual distance adjustment of multidimensional data points has been proposed when mapping them nonlinearly to a two-dimensional space, which improves the visualization quality, better exposes data clusters, and less distorts the structure of multidimensional data.

4. Two parameterization systems of physiological data have been explored and compared, namely: the parameterization system of fractal dimensions and the polynomial approximation parameter system. Visualizing the data of fractal dimensions, we notice explicit overlapping of groups. The best classification result was achieved by classifying capacity dimension data. After classifying these data, we have established that SVM classifier is the best one (gen. accuracy is $\approx 90\ \%$, sensitivity $\approx 80\ \%$). When visualizing data of the polynomial parameter system, overlapping of groups is not so notable. After classifying these data, we see that the SVM classifier is also most precise (gen. accuracy is $\approx 92\ \%$, sensitivity $\approx 83\ \%$).

5. Relying on the analysis of the polynomial parameter system, based on the correlation and visual analysis, we have defined that the parameter group of heart rate (HR) is strongly dependent on the JT interval parameter group, thus, we can refuse from the HR parameters group.

6. The research leads to the following summarised scheme for preliminary evaluation of the human health state:

   (1) Development of the visual classifier: (a) selecting a parameter system, that describes a patient's health state and that is directly derived from the physiological features, (b) evaluation of these parameters for a sufficiently large set of patients, (c) data analysis (classification, clustering, and visualization) of the patients including the known doctor's diagnosis.
   (2) Preliminary diagnosis of a new patient: mapping of the point, corresponding to this patient, among the previously mapped points; visual evaluation of the position of the mapped point among the decision surfaces.
   The proposed method is useful for medics while valuating the health state of patients and observing health worsening of a sportsman at the initial stage.

7. The methodology proposed was applied in the analysis of physiological data, but it can also be applied in any analysis of medical data with a view to make a preliminary diagnosis as well as to analyze multidimensional data of a general nature. However, in the latter case, it is necessary to go deeper into the origin and specificity of data.

List of published works on the topic of the dissertation

Article in scientific publications from the Thomson ISI Web of Science list

1. Bernatavičienė, J.; Dzemyda, G.; Marcinkevičius, V. 2007. Conditions for optimal efficiency of relative MDS, *Informatica* 18(2): 187–202. ISSN 0868-4952.

Articles in the scientific publications from the Thomson ISI Proceedings list

2. Bernatavičienė, J.; Šaltenis, V. 2006. Correction of distances in the visualization of multidimensional data, *Series on computers and operations research 7, Computer aided methods in optimal design and operations,* New Jersey. [etc.]: World Scientific, 159–168. ISBN 981-256-909-X.
3. Bernatavičienė, J.; Dzemyda, G.; Kurasova, O.; Marcinkevičius, V. 2006. Decision support for preliminary medical diagnosis integrating the data mining methods, *Simulation and optimisation in business and industry:* International conference on operational research: May 17–20, 2006, Kaunas: Technologija, 155–160. ISBN 9955-25-061-5.

Articles in periodical scientific publications from the list approved by the Science Council of Lithuania

4. Bernatavičienė, J.; Dzemyda, G.; Kurasova, O.; Marcinkevičius, V. 2006. Strategies of selecting the basic vector set in the relative MDS, *Ūkio technologinis ir ekonominis vystymas [Technological and economic development of economy]* 12(4): 283–288. ISSN 1392-8619.
5. Bernatavičienė, J.; Dzemyda, G.; Kurasova, O.; Marcinkevičius, V.; Medvedev, V. 2007. The Problem of Visual Analysis of Multidimensional Medical Data. *Models and Algorithms for Global Optimization, Springer Optimization and Its Applications* 4: 277–298. ISSN 1931-6828 [SpringerLINK].

Articles in the proceedings of scientific conferences

6. Bernatavičienė, J.; Dzemyda, G.; Kurasova, O.; Marcinkevičius, V.; Šaltenis, V.; Tiešis, V. 2006. Visualization and analysis of the eye fundus parameters, in *Proceedings of the 6th Nordic conference on eHealth and telemedicine NCeHT2006*, Finland, Helsinki, 267–268.
7. Bernatavičienė, J.; Šaltenis, V. 2005. Atstumų koregavimas vizualizuojant daugiamačius duomenis, *Informacinės technologijos 2005: konferencijos pranešimų medžiaga*, Kaunas, 102–107. ISBN 9955-09-788-4.
8. Bernatavičienė, J.; Berškienė, K.; Ašeriškytė, D.; Dzemyda, G.; Vainoras, A.; Navickas, Z. 2005. Fraktalinių dimensijų biomedicininio

informatyvumo analizė, *Biomedicininė inžinerija: tarpt. konferencijos pranešimų medžiaga*: Kaunas, 27–31. ISBN 9955-09-950-X.

9. Bernatavičienė, J.; Dzemyda, G.; Kurasova, O.; Vainoras, A. 2006. Integration of classification and visualization for diagnosis decisions, *International journal of information technology and intelligent computing* 1(1): 57–68. ISSN 1895-8648.

Short description about the author of the dissertation

1991–2006 – Studies at the Vilnius Pedagogical University, Faculty of Mathematics and Informatics – Bachelor of Mathematics. 2002–2004 – Studies at the Vilnius Pedagogical University, Faculty of Mathematics and Informatics – Master of Informatics. 2004–2008 – PhD studies at the Institute of Mathematics and Informatics, Systems Analysis Department.

VIZUALIOS ŽINIŲ GAVYBOS METODOLOGIJA IR JOS TYRIMAS

Tyrimų sritis ir problemos aktualumas. Duomenų suvokimas yra ilgo ir sudėtingo duomenų analizės proceso rezultatas. Jis apima daug etapų: suformuluojami analizės tikslai ir uždaviniai; iškeliamos pirminės hipotezės apie duomenų struktūras; formuojama duomenų imtis tyrimams; pasirenkami, kuriami nauji duomenų gavybos ir analizės metodai; analizuojami duomenų gavybos metodais gauti rezultatai; gautos žinios apibendrinamos, paneigiamos arba priimamos iškeltos hipotezės. Šio darbo tyrimų sritis yra žinių gavybos iš daugiamačių duomenų procesas ir tiriamų duomenų suvokimo gerinimo būdai.

Technikoje, medicinoje, ekonomikoje, ekologijoje ir daugelyje kitų sričių nuolat susiduriama su daugiamačiais duomenimis. Vystantis technologijoms, tobulėjant kompiuteriams ir programinei įrangai, kaupiamų duomenų apimtys ypač sparčiai didėja. Tačiau tebelieka didelė spraga tarp duomenų surinkimo bei saugojimo, jų suvokimo bei gautų žinių pritaikymo sprendžiant praktinius uždavinius. Daugiamačių duomenų suvokimas yra rezultatas ilgo ir sudėtingo duomenų analizės proceso. Šis procesas – tai perėjimas nuo didelės analizuojamos duomenų aibės prie specifinių duomenų, iš kurių išskiriama informacija bei suformuojamos žinios apie tiriamų duomenų struktūrą, naujus sąryšius, duomenų grupes, kas turės įtaką tolimesnių sprendimų priėmimui.

Atskiri duomenų analizės proceso etapai yra detaliai išnagrinėti literatūroje, tačiau trūksta vientisos, visus duomenų analizės etapus apimančios, metodologijos. Ji įgalins tyrėją iš turimų duomenų išgauti maksimalų informacijos kiekį, apjungti šią informaciją su eksperto patirtimi ir suformuoti žinių banką, kuris padės išspręsti tyrime iškeltus uždavinius.

Sprendžiama **problema** – vizualaus žinių gavybos proceso vientisumo užtikrinimas.

Darbo tikslas ir uždaviniai. Pagrindinis disertacijos tikslas yra sukurti ir ištirti vizualios žinių gavybos metodologiją, kuri leistų padidinti duomenų analizės efektyvumą. Norint pasiekti šį tikslą, reikėjo išspręsti tokius uždavinius: (1) analitiškai apžvelgti duomenų gavybos ir analizės metodus: klasifikavimo, klasterizavimo ir vizualizavimo; (2) išanalizuoti žinių gavybos procesą, apžvelgti ir palyginti esamus šio proceso modelius, ištirti vizualizavimo galimybių panaudojimą žinių gavybos procese; pasiūlyti ir ištirti daugiamačių duomenų vizualizavimo proceso modelį, kuriuo pagrindžiama kuriama metodologija; (3) ištirti pasirinktus algoritmus, naudojamus vizualios duomenų gavybos procese, ir sukurti efektyvesnes jų modifikacijas; ištirti naujų (papildomai gautų) daugiamačių duomenų vizualizavimo galimybes bei pagerinti tam naudojamų metodų efektyvumą; (4) pasiūlyti ir ištirti daugiamačių duomenų išdėstymo geometrijos keitimo būdus, siekiant tikslesnės analizuojamų duomenų projekcijos plokštumoje; (5) pritaikyti sukurtą metodologiją medicininių ir fiziologinių duomenų analizei.

Tyrimų **metodikos** pagrindą sudaro analitinė analizė, apibendrinimas ir eksperimentinis tyrimas.

Tyrimų objektas. Disertacijos tyrimų objektas – vizualios žinių gavybos procesas. Su šiuo objektu betarpiškai susiję dalykai: (1) daugiamačių duomenų pirminės aibės suformavimas; (2) klasterizavimo, vizualizavimo ir klasifikavimo algoritmai; (3) duomenų gavybos metodais gautų rezultatų įvertinimas; (4) naujų daugiamačių duomenų atvaizdavimas; (5) sprendimų priėmimas ir gautų žinių apibendrinimas, atsižvelgiant į analizės rezultatus.

Mokslinis naujumas. Sukurta vizualios žinių gavybos metodologija, kuri leidžia atlikti išsamią ir informatyvią tiriamų duomenų analizę. Pasiūlyti santykinių daugiamačių skalių metodo efektyvumo gerinimo būdai: (1) sukurtos bazinių vektorių parinkimo strategijos; (2) ištirtos inicializavimo problemos santykinių DS algoritme, nustatytas geriausias dvimačių vektorių inicializavimo būdas; (3) pasiūlytas optimalaus bazinių vektorių skaičiaus parinkimo būdas. Sukurtas atstumų tarp daugiamačių duomenų koregavimo algoritmas, kuris pagerina vizualizavimo kokybę: geriau išryškina duomenų klasterius, mažiau iškraipo daugiamačių duomenų struktūras; (4) pasiūlytas preliminaraus sveikatos būklės fiziologinių duomenų analizės pagrindu įvertinimo būdas, besiremiantis sukurta metodologija.

Ginamieji teiginiai

1. Vizualios žinių gavybos proceso susisteminimas leidžia visapusiškai įvertinti ir pritaikyti vizualizavimo metodų ir priemonių teikiamas galimybes duomenų analizės efektyvumui didinti.
2. Santykinių daugiamačių skalių efektyvumą galima pagerinti tinkamai parenkant bazinių vektorių skaičių, bazinių vektorių parinkimo strategiją bei dvimačių vektorių inicializavimo būdą.
3. Daugiamačių vektorių vizualizavimo kokybę galima pagerinti taikant daugiamačių duomenų atstumų koregavimo transformaciją.
4. Sukurtą vizualios žinių gavybos metodologiją galima taikyti preliminariam sveikatos būklės vertinimui.

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: 1) 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.; 2) Lietuvos valstybinio mokslo ir studijų aukštųjų technologijų plėtros programos projektą „Informacinės klinikinių sprendimų palaikymo ir gyventojų sveikatinimo priemonės e. Sveikatos sistemai (Info Sveikata)“; Registracijos Nr.: B-07019; Vykdymo laikas: nuo 2007 m. 09 mėn.

Darbo rezultatų aprobavimas ir publikavimas. Tyrimų rezultatai publikuoti 9 moksliniuose leidiniuose: 1 straipsnis leidinyje, įtrauktame į Mokslinės informacijos instituto pagrindinį (Thomson ISI Web of Science) sąrašą; 2 straipsniai leidiniuose, įtrauktuose į Mokslinės informacijos instituto konferencijos darbų (Thomson ISI Proceedings) duomenų bazę; 2 straipsniai Lietuvos mokslo tarybos patvirtinto sąrašo tarptautinėse duomenų bazėse referuojamuose leidiniuose; 1 straipsnis recenzuojamoje konferencijų pranešimų medžiagoje ir 3 straipsniai kituose periodiniuose bei vienkartiniuose straipsnių rinkiniuose.

Tyrimų rezultatai buvo pristatyti ir aptarti devyniose konferencijose.

Darbo apimtis. Disertaciją sudaro penki skyriai ir literatūros sąrašas. Bendra disertacijos apimtis 116 puslapių, 44 paveikslai ir 12 lentelių.

Pirmajame 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.

Antrajame skyriuje yra atlikta duomenų gavybos sprendžiamų uždavinių ir metodų šiems uždaviniams spręsti analitinė apžvalga. Išanalizuotas duomenų gavybos ir analizės procesas, apžvelgti ir palyginti keturi šio proceso modeliai. Parodyta, kad pagrindinis jų skirtumas yra detalizacijos lygmenyje. Apibendrinus gauta šešių žingsnių schema, kuri apima visus žinių radimo etapus, įvardinama aiški vizualizavimo vieta kiekviename žinių radimo etape. Analizė parodė, kad vizualizavimas užima svarbią vietą duomenų gavybos ir analizės procese, tačiau jo panaudojimas yra gana fragmentiškas. Čia yra reikalingas kompleksiškas požiūris apimantis visus analizės proceso etapus.

Trečiajame skyriuje pateikiama vizualios žinių gavybos metodologija, santykinių daugiamačių skalių metodo efektyvumo gerinimo būdai bei daugiamačių duomenų tarpusavio atstumų koregavimo algoritmas, naudojamas vizualizuojant daugiamačius duomenis.

Ketvirtajame skyriuje pateikta fiziologinių duomenų analizė, naudojant vizualios žinių gavybos metodologiją.

Penktajame skyriuje pateiktos disertacijos išvados.

Bendrosios išvados ir rekomendacijos

1. Vizualios žinių gavybos proceso susisteminimas leidžia visapusiškai įvertinti ir pritaikyti vizualizavimo metodų ir priemonių teikiamas galimybes duomenų analizės efektyvumui didinti.

2. Detaliai ištyrus santykinių DS algoritmą, galime daryti šias išvadas:

   - Vizualizavimo rezultatai priklauso nuo bazinių vektorių parinkimo strategijos (kuo tolygiau baziniai vektoriai pasiskirstę po visą tiriamą aibę, tuo tikslesnė projekcija yra gaunama) ir bazinių vektorių skaičiaus, dvimačių vektorių inicializavimo būdo.
   - Naudojant inicializavimo būdą, paremtą PCA algoritmu, paklaidos vidurkis mažesnis už paklaidų vidurkius gaunamus kitomis strategijomis, tačiau skirtumai tarp šių vidurkių nereikšminiai. Blogiausias inicializavimo būdas yra atsitiktinis taškų parinkimas bazinių vektorių projekcijų srityje.
   - Vizualizuojant daugiamačius duomenis, kai tiriamų duomenų dimensija yra didesnė už 5, o duomenų aibę sudaro daugiau nei 3000 vektorių, tikslingiau vietoj standartinio daugiamačių skalių algoritmo naudoti santykinių DS algoritmą. Didinant bazinių vektorių skaičių gaunama tikslesnė projekcija. Tačiau per didelis bazinių vektorių skaičius lėtina skaičiavimus. Tyrimai parodė, kad mažesnėms duomenų aibėms (iki

3000 vektorių) tikslinga imti nuo 700 iki 1000 bazinių vektorių, o didelėms duomenų aibėms – nuo 900 iki 1500.

3. Pasiūlyta daugiamačių duomenų taškų tarpusavio atstumų koregavimo transformacija, atliekant jų netiesinį projektavimą į dvimatę plokštumą, pagerina vizualizavimo kokybę, koregavimas geriau išryškina duomenų klasterius, mažiau iškraipo daugiamačių duomenų struktūras.

4. Buvo ištirtos ir palygintos dvi fiziologinių duomenų parametrizavimo sistemos: fraktalinių dimensijų parametrizavimo sistema ir polinominio aproksimavimo parametrų sistema. Vizualizuojant fraktalinės dimensijos duomenis matomas grupių persidengimas. Geriausias klasifikavimo rezultatas gautas klasifikuojant užimtumo dimensijos duomenis. Atlikus šių duomenų klasifikavimą nustatyta, kad tiksliausiai klasifikuoja atraminių vektorių (SVM) klasifikatorius (bendras tikslumas $\approx 90\,\%$, jautrumas $\approx 80\,\%$). Vizualizuojant polinominės parametrų sistemos duomenis, grupių persidengimas mažesnis. Atlikus šių duomenų klasifikavimą, tiksliausias yra taip pat atraminių vektorių klasifikatorius (bendras tikslumas $\approx 92\,\%$, jautrumas $\approx 83\,\%$).

5 Remiantis atlikta polinominės parametrų sistemos analize, kuri paremta koreliacine bei vizualia analize, nustatyta, kad dydžio ŠSD parametrų grupė yra labai priklausoma nuo dydžio JT parametrų grupės, ir ŠSD parametrų grupės galima atsisakyti.

6. Norint nustatyti preliminarią diagnozę naujam pacientui, yra siūloma rasti paciento duomenų projekciją plokštumoje, kur jau yra fiksuota etaloninė bazinių vektorių projekcija (suprojektuoti tiriamųjų duomenys su jau nustatytomis diagnozėmis, nubrėžti klasių skiriamieji paviršiai) ir nustatyti naujo taško padėtį tarp esamų taškų. Priklausomai nuo taško, atitinkančio tiriamąjį pacientą, padėties tarp skiriamųjų paviršių galime daryti preliminarų sprendimą apie šio tiriamojo sveikatos būklę. Pasiūlytas būdas yra naudingas medikams vertinant pacientų sveikatos būklę ir pastebint sportininko sveikatos pablogėjimą pradinėje to stadijoje.

7. Siūloma metodologija buvo taikyta fiziologinių duomenų analizei, tačiau ją galima taikyti bet kokių medicininių duomenų analizei siekiant nustatyti preliminarią diagnozę, o taip pat bendro pobūdžio daugiamačiams duomenims analizuoti. Tačiau pastaruoju atveju būtina įsigilinti į tų duomenų kilmę ir specifiką.

Jolita Bernatavičienė

METHODOLOGY OF VISUAL KNOWLEDGE DISCOVERY AND ITS INVESTIGATION

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

Jolita Bernatavičienė

VIZUALIOS ŽINIŲ GAVYBOS METODOLOGIJA IR JOS TYRIMAS

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

2008 05 09. 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, *http://www.kopija.lt*