VYTAUTAS MAGNUS UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Rasa KARBAUSKAITĖ

ANALYSIS OF MULTIDIMENSIONAL DATA VISUALIZATION METHODS THAT PRESERVE THE LOCAL STRUCTURE

Summary of Doctoral Dissertation

Physical Sciences (P 000)
Informatics (09 P)
Informatics, Systems Theory (P 175)

Vilnius, 2010

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2006–2010.

Scientific Supervisor

Prof Dr Habil Gintautas DZEMYDA (Institute of Mathematics and Informatics, Physical Sciences, Informatics – 09 P).

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

Chairman

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

Members:

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

Doc Dr Rimantas VAICEKAUSKAS (Vilnius University, Physical Sciences, Informatics – 09 P),

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

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

Opponents:

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

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 in the auditorium number 203 of the Institute of Mathematics and Informatics at 10 a.m. on September 28 2010.
Address: Akademijos str. 4, LT-08663 Vilnius, Lithuania.

The summary of the dissertation was distributed on 27 August, 2010.
A copy of the doctoral dissertation is available for review at the M. Mažvydas National Library of Lithuania, the Library of the Vytautas Magnus University and at the Library of the Institute of Mathematics and Informatics.

VYTAUTO DIDŽIOJO UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Rasa KARBAUSKAITĖ

DAUGIAMAČIŲ DUOMENŲ VIZUALIZAVIMO METODŲ, IŠLAIKANČIŲ LOKALIĄ STRUKTŪRĄ, ANALIZĖ

Daktaro disertacijos santrauka

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

Vilnius, 2010

Disertacija rengta 2006–2010 metais Matematikos ir informatikos institute.

Mokslinis vadovas

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

Disertacija ginama Vytauto Didžiojo universiteto Informatikos mokslo krypties taryboje:

Pirmininkas

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

Nariai:

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

doc. dr. Rimantas VAICEKAUSKAS (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P),

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

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

Oponentai:

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

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

Disertacija bus ginama viešame Informatikos mokslo krypties tarybos posėdyje 2010 m. rugsėjo mėn. 28 d. 10 val. Matematikos ir informatikos institute, 203 auditorijoje. Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2010 m. rugpjūčio 27 d.
Disertaciją galima peržiūrėti M. Mažvydo nacionalinėje bibliotekoje, Vytauto Didžiojo universiteto ir Matematikos ir informatikos instituto bibliotekose.

1. Introduction

Relevance of the Problem

There is none activity field of mankind that would not collect and analyse multidimensional data. A typical example of multidimensional data is related with image processing. Frequently data are comprised of photos of the same object, obtained by gradually turning the object at a certain angle or by taking its photos at different moments. Each photo is digitized, i.e., the coordinates of a data point consist of colour parameters of photo points and therefore the number of the coordinates of this point is very large. Dimensionality reduction methods (projection methods) are developed rather intensively. By transforming multidimensional data into a two- or three-dimensional space and after visualizing them, it is much easier to conceive the structure of data and connections among them. However, while transforming data into a lower dimensional space, data projection distortions are inevitable. That is why evaluation of the projection quality obtained remains a topical problem.

We often work with datasets that are constantly supplemented with new data. It is of great importance to immediately map the new data points without loss of a high accuracy. Therefore, the mapping of new points, their insertion among the earlier mapped points is one of the problems considered in the thesis.

Projection methods of multidimensional data come across two main problems. The first one is to find multidimensional data projections in a space of lower dimensionality (two or three-dimensional space) with a view to preserve the proximity (similarities or dissimilarities) of objects of the analysed set as exactly as possible. The second one is to map multidimensional data in a low-dimensional space so that their projections did not overlap. This problem is also one of the problems solved in the thesis.

Frequently in practical problems multidimensional data are accumulated and the points corresponding to them are considered in a high-dimensional space, while in fact they are either points of a manifold of some lower dimensionality or the points close to that manifold. Thus, one of the major problems of the thesis is to discover a low-dimensional nonlinear manifold in a high-dimensional data space and then transfer the data points that lie on or near to this manifold, into this low-dimensional space.

An important point of a manifold is its topology. There are a lot of different measures of topology preservation in the literature. Another important problem solved in the thesis is to find and explore those measures that would be suitable to analyse the manifold topology preservation after its transformation into a low-dimensional space.

The Object of Research

The research object of this work is multidimensional data visualization algorithms and methods that preserve the local structure of multidimensional data, as well as the evaluation criteria of multidimensional data projections in a low-dimensional space. Here the preservation of a local structure means a preservation of distances among the nearest points after transforming the multidimensional dataset analysed from a high-dimensional space to a low-dimensional one.

The Aim and Tasks of the Thesis

The key aim of the work is to analyse multidimensional data visualization algorithms that preserve the local structure of the analysed data, to modify them as well as to investigate the importance of control parameters of the algorithms analysed and to propose ways of selecting the parameters in order to get a data projection as precise as possible.

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

1. To analyse the existing visualization methods of multidimensional data and to determine the group of the analysed visualization methods that preserve a local structure;
2. To analyse the chosen methods and visualize multidimensional data by them;
3. To compare the obtained visualization results obtained by the investigated methods with the results obtained by the methods that preserve not only the local structure;
4. To estimate the quality of data projections obtained by various visualization methods (using expert and quantitative numerical measures);
5. To modify multidimensional data visualization algorithms that preserve the local structure in order to get a projection of the data analysed as precise as possible;
6. To evaluate the results of the modified algorithms in comparison with that obtained by the original algorithms;
7. To investigate the importance of control parameters of the analysed algorithms and to propose the ways of selecting these parameters in order to get a projection of the analysed data as precise as possible.

Scientific Novelty

1. Realizations of the triangulation method have been explored experimentally, using the method of the second nearest neighbour and that of the reference point to select the reference points. It has been established that, in both cases, the projection error strongly depends on the order of points to be mapped, which proves that it does not suffice to use the triangulation method alone for data visualization. A new realization of the combination of Sammon and triangulation methods has been proposed for visualizing the new points.
2. The dependence of the relational perspective map (RPM) on the parameters (the width and height of the rectangle, in which data points are visualized, and the initial learning speed) has been investigated experimentally. A new realization of the RPM method has been proposed, which allowed us to nearly avoid this dependence.
3. The dependence of locally linear embedding (LLE) on the parameters – the number of the nearest neighbours of each data point and the regularization parameter of a local Gram matrix – has been investigated experimentally. With a view to obtain the most precise data projections possible, a new way for selecting the number of the nearest neighbours in the LLE algorithm has been proposed. A new regularization algorithm of the local Gram matrix has been created as well.
4. The importance of parameters (the number of the nearest neighbours of each data point and the heat kernel parameter that is used in the Gaussian kernel function while computing the weights) in the Laplacian Eigenmaps (LE) algorithm has been explored experimentally. A modification of the LE algorithm has been proposed, where only one significant control parameter – the maximal number of the nearest neighbours – remains.
5. Three topology preservation measures, Spearman's rho, Konig's measure (KM), and mean relative rank errors (MRRE), that are suitable to analyse the topology preservation of a manifold after its transformation to a low-dimensional space, have been investigated and compared using two criteria: the topology preservation quality and computational expenditure.

The Defended Statements

1. Saving computational expenditure (time) and losing accuracy not so much, the triangulation method can be used for visualizing the new multidimensional points, if the initial data points were projected by the MDS-type method.

2. The control parameters of the relational perspective map, locally linear embedding, and Laplacian Eigenmaps methods greatly influence the quality of the projections obtained. However, there are some strategies so that it is possible to decrease the number of parameters or regulate the selection of their values.
3. Konig's measure (KM) and mean relative rank errors (MRRE) always define well the topology preservation of a manifold after its transformation to a low-dimensional space, while Spearman's rho can only be successfully applied to estimate the topology preservation of manifolds of a simpler structure.

Practical Significance

The results of investigations revealed possibilities for analysing manifold-type multidimensional data. It has been shown that nonlinear manifold learning methods can be widely used in various areas including medicine, too.

The research was partly supported by the Lithuanian State Science and Studies Foundation project ”Information technology tools of clinical decision support and citizens wellness for e.Health system (No. B-07019)” Start date: 09-2007; finish date: 12-2009.

Approbation and Publications of the Research

The main results of this dissertation were published in 6 scientific publications: 2 articles in international periodicals included into ISI Web of Science; 2 articles in other scientific journals reviewed; 2 articles in the proceedings of scientific conferences. The full list of publications on the topic of the dissertation is presented at the end of this Summary. The main results of the work have been presented and discussed at 5 national and international conferences.

The Scope of the Scientific Work

The work is written in Lithuanian. It consists of 8 chapters and the list of references. There are 168 pages of the text, 103 figures, 6 tables, and 123 bibliographical sources in the thesis.

2. Analysis of Dimensionality Reduction Methods of Multidimensional Data

In Section 2, an analytic review of dimensionality reduction methods of multidimensional data is made. We have systematized and explored those dimensionality reduction methods that preserve only a local structure, i.e., distances among the nearest data points, while transforming multidimensional data to a low-dimensional space. Such methods are: triangulation, relational perspective map, locally linear embedding, Laplacian Eigenmaps, Hessian eigenmaps, and local tangent space alignment. We have also briefly reviewed frequently used dimensionality reduction methods (principal component analysis, multidimensional scaling) that try to preserve not only a local, but also a global structure, i.e., distances among all the data points, because the drawbacks of these methods, revealed while investigating certain datasets (manifold type data), highlighted the advantages of the methods analysed.

3. Investigation of the Triangulation and Sammon Methods and Their Combination

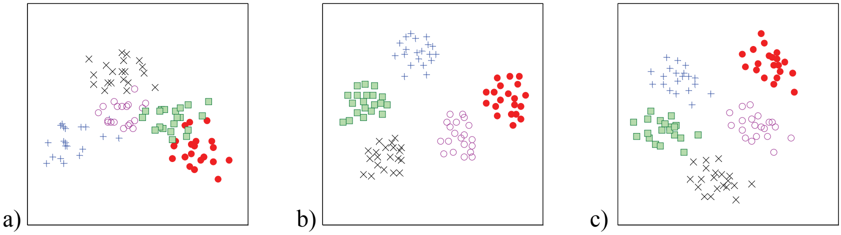
In Section 3, two nonlinear dimensionality reduction methods of multidimensional data are investigated: Sammon's projection and the triangulation method as well as their combination. The Sammon projection belongs to the methods of the multidimensional scaling group (MDS). Sammon's method is a method of simultaneous mapping (all data points are mapped at once), and we try to preserve relative distances among all data points by this method. Triangulation is a method of sequential mapping (data points are mapped not all at once, but sequentially one point after another; a point is mapped with regard to the position of the earlier mapped points). Having mapped a new point by the triangulation method, its distances are exactly preserved only to two points previously mapped; the distance between the two points is preserved exactly as well.

The triangulation method is fast enough, however, it can preserve only $(2m-3)$ distances among the points analysed (m is the number of points). Therefore the projection error E_S is rather big (Fig. 1a). Besides, it has been established that the projection error greatly depends on the order of the points to be mapped. The Sammon algorithm tries to preserve all $m(m-1)/2$ distances among data points, so the projection error E_S is not big, but the algorithm is rather slow (Fig. 1b). A new realization of the combination of Sammon and triangulation methods is proposed in this section. It enables us to map the new multidimensional points rather precisely and fast without Sammon mapping of the whole (updated) dataset (Fig. 1c).

The projection error (Sammon‘s stress (error)) is calculated by the formula

$$E_S = \sum_{\substack{i,j=1\\ i<j}}^{m} \left(\left(d(X_i,X_j) - d(Y_i,Y_j) \right)^2 / d(X_i,X_j) \right) \Big/ \sum_{\substack{i,j=1\\ i<j}}^{m} d(X_i,X_j),$$

where $d(X_i,X_j)$ and $d(Y_i,Y_j)$ are distances, respectively, between the multidimensional data points X_i and X_j and two-dimensional points Y_i and Y_j corresponding to them.



$t = 683$ ms, $E_S = 0{,}1821811$

$t = 5125$ ms, $E_S = 0{,}0448481$

$t = 2952$ ms, $E_S = 0{,}0462552$

Fig. 1. Visualization of the test data: a) 100 points, mapped by the triangulation method (75 initial, 25 new points), b) 100 points, mapped by Sammon’s method, c) 75 initial points, mapped by Sammon’s method, and 25 new points mapped by triangulation

4. Investigation of Realizations of the Relational Perspective Map

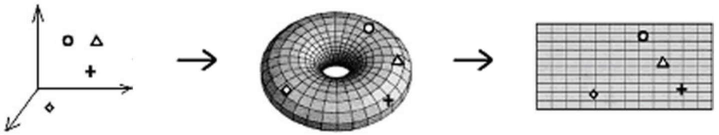
Relational perspective map (RPM) is a dimensionality reduction method which is intended to visualize multidimensional data on the surface of a torus (in a rectangle). RPM, like MDS methods, tries to preserve the relational distances among the multidimensional data in a low-dimensional space. But the most important feature of the relational perspective map is the ability to partition complex datasets into less complex pieces and then map data points on the torus surface in a non-overlapping manner. Therefore, it preserves short-range distances better than other projection methods.

In order to visualize data points in a low-dimensional space in a non-overlapping manner while preserving relational distances of the original data points as much as possible, the surface of a torus is chosen for data mapping. From the physical point of view, a torus is a force-directed multiparticle system: the projection points are considered as particles that can move freely on the surface of the torus, but cannot escape the surface. The particles exert repulsive forces on one another so that, guided by the forces, the particles rearrange themselves to a configuration that visualizes the relational distances among the original data points. While mapping data points on a torus, the RPM algorithm minimizes the potential energy:

$$E_p = \sum_{i<j} d\left(X_i, X_j\right)/pd^p\left(Y_i, Y_j\right), \text{ and } E_0 = -\sum_{i<j} d\left(X_i, X_j\right) ln\left(d\left(Y_i, Y_j\right)\right).$$

Here the parameter $p \in (-1; +\infty)$ is called the rigidity, $d\left(X_i, X_j\right)$ and $d\left(Y_i, Y_j\right)$ are distances, respectively, between the original data points X_i and X_j and two-dimensional points Y_i and Y_j corresponding to them.

As illustrated in Fig. 2, the RPM algorithm first maps data points onto the surface of a torus. Then by cutting the torus vertically at a certain place and afterwards horizontally through a generatrix of the obtained cylinder, we obtain a rectangle in which we see the projections of multidimensional data.



A dataset with 4 points in a three-dimensional space

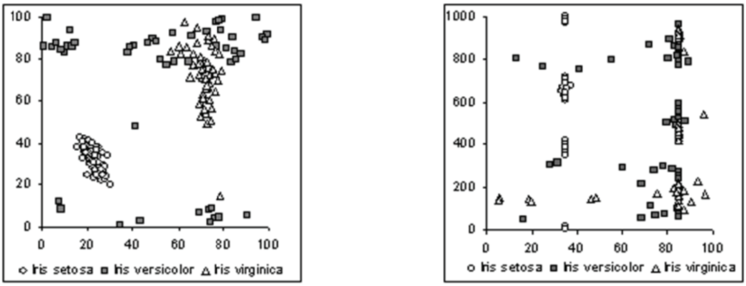
4 two-dimensional points on the torus surface

RPM mapping of the dataset

Fig. 2. The model of the RPM method

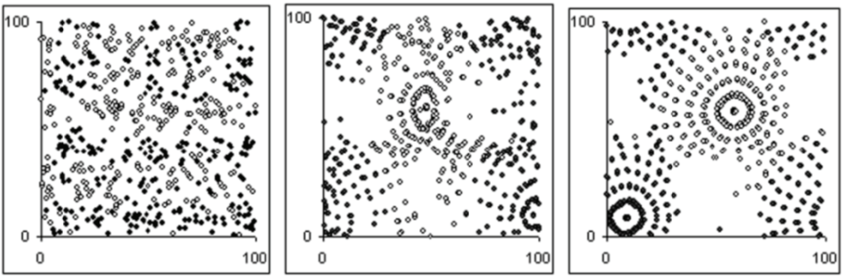
In this section, the RPM algorithm is explored in detail theoretically and experimentally in order to find out how much the obtained mapping of multidimensional points depend on the values of the parameters $\tilde{r}$ (the initial learning speed), w (the width of the rectangle) and h (the height of the rectangle). Figures 3 and 4 show that the values of these parameters greatly influence the data projections obtained. When pursuing investigations with the spherical dataset, it has been noticed: though the relative value of the potential energy changes fractionally between $(\tilde{r} = 0{,}5)$ and $(\tilde{r} = 4)$, i.e., only 4%, however, even an insignificant reduction of the potential energy substantially changes the projections of points (Fig. 4).

Since there are no specific rules to select the values of these parameters, a new RPM modification is proposed in this thesis. In our modification, we reject the parameter $\tilde{r}$ and avoid a strong dependence on the parameters w and h. As a result, mappings on a plane are very similar (in terms of distinguishing the classes) (Fig. 5).



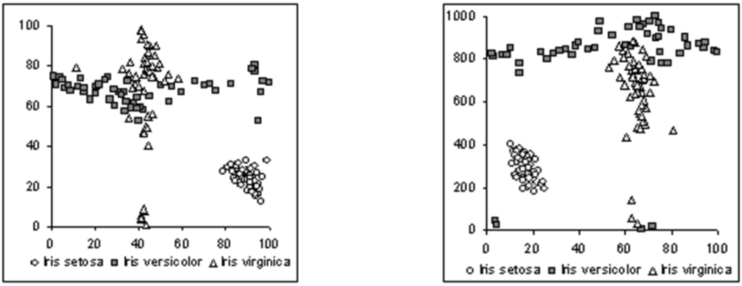
a) $w = 100, h = 100$ b) $w = 100, h = 1000$

Fig. 3. Visualization of the iris dataset on the plane (in the rectangle) by the RPM method at different values of the parameters w and h ($\tilde{r} = 4$)



$\tilde{r} = 0{,}5, E_p = -3342223$ $\tilde{r} = 2, E_p = -3472738$ $\tilde{r} = 4, E_p = -3491417$

Fig. 4. Visualization of the spherical dataset on the plane (in the rectangle) by the RPM method at different values of the parameter $\tilde{r}$ ($w = h = 100$)



a) $w = 100, h = 100$ b) $w = 100, h = 1000$

Fig. 5. Visualization of the iris dataset on the plane (in the rectangle) by modifying the RPM algorithm at different values of the parameters w and h

5. Strategies for Selecting Parameters of Locally Linear Embedding

Nonlinear manifold learning methods are based on an assumption that data points lie on a low-dimensional nonlinear manifold embedded in a high-dimensional space. The key purpose of these methods is to discover the low-dimensional nonlinear manifold in a high-dimensional data space and then transfer the data points that lie on or near to this manifold into a low-dimensional space, preserving the underlying structure in the data.

In the second section, after an exhaustive analysis of nonlinear manifold learning methods and a proper review of their application areas, we can conclude: these methods are very popular at present and widely applicable, especially in image processing. Therefore it is reasonable to concentrate further investigations on nonlinear manifold learning methods, with a view to increase the efficiency of these methods even more.

In the fifth section, one of the nonlinear manifold learning methods – locally linear embedding (LLE) – is explored in detail. The properties of this method are: while mapping multidimensional data to a low-dimensional space, only the relationships among the nearest points are preserved; the global structure of a nonlinear manifold (overall arrangement of points of the dataset, considering the existence of a low-dimensional nonlinear manifold) is discovered; data unambiguously are mapped into a low-dimensional space.

Let us analyse 1000 three-dimensional data points that lie on a nonlinear two-dimensional S-manifold (Fig. 6a). Projections of these points onto a plane, obtained by LLE (Fig. 6b) and MDS (Fig. 6c.) methods, are shown in Fig 6. The dimensionality reduction by LLE succeeds in identifying the underlying structure of the manifold, while the MDS method maps faraway data points on the manifold to nearby points on the plane, failing to identify the structure.

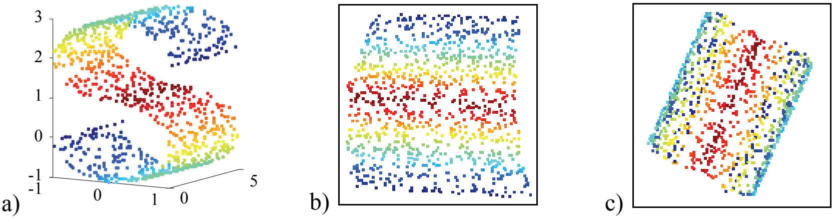


Fig. 6. a) 1000 3-dimensional data points on an S-manifold, b) projection of these points on a plane, obtained by LLE, c) projection of these points on a plane, obtained by MDS

The LLE algorithm has three control parameters: number k of the nearest neighbours of each data point, the regularization parameter ε of a local Gram matrix and the intrinsic dimensionality d of the data (dimensionality d of a manifold). The data mapping quality is strongly dependent on these parameters.

In this section, we propose new ways for selecting the values of the parameters k and ε. When solving a manifold learning problem, data points must be transferred into a space of such a dimensionality that is the dimensionality d of the manifold. We choose $d = 2$, because we analyse only those data points that lie on the two-dimensional manifolds in this thesis.

Selection of the number k of the nearest neighbours in the LLE algorithm. The experiments have shown that the values of this parameter greatly influence the results obtained: if k is set too small, the continuous manifold can falsely be divided into disjoint sub-manifolds and the mapping does not reflect any global properties; if k is too high, a large number of the nearest neighbours causes smoothing or elimination of small-scale structures in the manifold, the mapping loses its nonlinear character and behaves like traditional PCA. Furthermore, the results of LLE are typically stable over a certain range of neighbourhood sizes. However, the size of that range depends on various features of the data, such as the sampling density and manifold geometry.

In this section, the proposed new way for selecting the number k of the nearest neighbours allowed us to define a proper range of neighbourhood sizes. The essence of this way is that the LLE algorithm is run with different values of the parameter k (the values of the parameter k are chosen from rather a wide interval) and each time, when projections are obtained, Spearman's rho is calculated. Then the dependence of Spearman's rho on k is drawn and an interval of the number of the nearest neighbours is found such that the values of Spearman's rho are maximal.

Spearman's rho is computed by using the following equation:

$$\rho_{Sp} = 1 - 6\sum_{i=1}^{S}\left(\hat{r}_X(i) - \hat{r}_Y(i)\right)^2 / (S^3 - S),$$

where S is the number of distances to be compared ($S = m(m-1)/2$), m is the number of data points, $\hat{r}_X(i)$, $i = \overline{1,S}$ are the ranks (order numbers) of pairwise distances calculated for the original (n-dimensional) data points and sorted in ascending order, and $\hat{r}_Y(i)$, $i = \overline{1,S}$ are the ranks (order numbers) of pairwise distances calculated for the projected (d-dimensional) data points and sorted in ascending order. As usual,$-1 \leq \rho_{Sp} \leq 1$. The best value of Spearman's rho is equal to one.

In the calculation of Spearman's rho, distances both on a plane and on a multidimensional space are used. A question arises which distances should be evaluated when estimating Spearman's rho: Euclidean or geodesic? The experiments have shown that the quantitative measure – Spearman's rho – is suitable to estimate the topology preservation after transferring the data to the two-dimensinal space by the LLE algorithm. In order that Spearman's rho properly reflected the projections obtained, when calculating its value it is necessary to evaluate the geodesic but not Euclidean distances in an n-dimensional space by selecting rather a small number of neighbours in the geodesic distance calculation algorithm ($k_{geod} \leq 10$). Since only two-dimensional manifolds were investigated, Euclidean distances were calculated on the plane.

We present here the investigation of the data points that lie on the S-manifold (Fig. 6a). Figure 7 shows three dependences of Spearman's rho on k: (I) Euclidean distances were evaluated in a space, (II) geodesic distances were evaluated in a space and a very small number of neighbours was fixed, i.e., $k_{geod} = 10$, when calculating these distances, (III) when calculating geodesic distances in a space, the number of neighbours is varying just like in the LLE algorithm, i.e., $k_{geod} = k$. In this section, it is shown that the structure of the S-manifold is destroyed if $k = 100$. Therefore, the values of dependences of Spearman's rho must be as low as possible for this value of the parameter k. Since the values of dependences (I) and (III) are close to 1 (respectively, ≈ 0.97 and ≈ 0.95) and the value of dependence (II) is lower (≈ 0.82), this investigation corroborates the fact that it is necessary to evaluate geodesic distances in an n-dimensional space by fixing rather a small number of neighbours for getting the values of these distances.

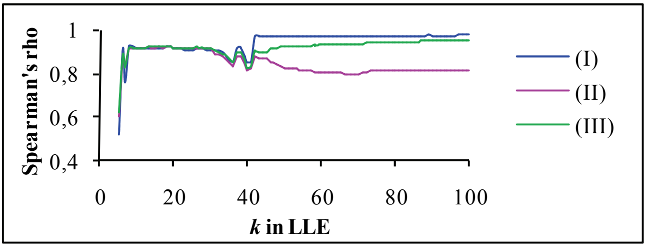


Fig. 7. Dependences of Spearman's rho on k obtained after transferring the points on the S-manifold to the plane by LLE: Euclidean distances were evaluated in a 3-dimensional space (I), and geodesic distances were evaluated in a 3-dimensional space, as $k_{geod} = 10$ (II), $k_{geod} = k$ (III)

Selection of the regularization parameter ε in the LLE algorithm. In the second step of the LLE algorithm, while looking for the optimal weights w_{ij}, it is necessary to calculate the inverse matrix of a local Gram matrix C^i (size $k \times k$) for each data point X_i. If the number of the nearest neighbours is larger than the dimensionality of the original data ($k > n$), then the Gram determinant $|C^i| = 0$, i.e., the matrix C^i is singular. However, due to calculation errors, $|C^i|$ may be not equal, but very close to zero, i.e., the Gram matrix is nearly singular. Hence, we cannot find the inverse matrix for the Gram matrix (if C^i is singular) or the weights w_{ij} are not uniquely defined (if C^i is nearly singular). In such cases, before looking for the inverse matrix for the Gram matrix, it is necessary to make a regularization of the Gram matrix C^i, i.e., the matrix C^i must be changed so that the determinant of the changed matrix be large enough.

In this section, a new regularization algorithm of the local Gram matrix is proposed: an explicit theoretical reasoning of the algorithm, its experimental investigation and comparison with the algorithm, proposed by Roweis and Saul (2000), are given.

Let us denote the regularization algorithm proposed by Roweis and Saul (2000) by R1, and the new one (proposed in this section) by R2. The essence of both regularization algorithms R1 and R2 is that we find the value of the regularization parameter ε, added to the elements of the main diagonal of matrix C^i and to the eigenvalues of this matrix. These algorithms differ only in control parameters that are chosen: in R1, the control parameter is a parameter t ($t > 0,\ t \ll 1$), and, in R2, it is a parameter D. However, an advantage of D over t is that D has a real sense (it is a determinant of the regularized Gram matrix), and t is only a certain multiplier. After choosing the values of these control parameters, the value of the regularization parameter ε is calculated by the formulas: $\varepsilon = \mathrm{tr}(C^i)t$, where $\mathrm{tr}(C^i)$ is a trace of C^i (in the case of R1) and $\varepsilon \approx \sqrt[k-r]{D/\prod_{j=1}^{r} \lambda_j}$, where r is the rank of matrix C^i, λ_j are eigenvalues of C^i (in the case of R2). In this section, it is shown that the control parameters D and t are closely related: by increasing the values of one parameter, the values the other one are increasing exponentially.

After investigating our regularization algorithm (R2) experimentally, it has been defined that the quality of the obtained projections strongly depends on the chosen values of the parameters D and k. The quality was evaluated by calculating Spearman's rho. It is very important to find the maximum value of Spearman's rho with the lower value of k, since the time of running the LLE algorithm increases in a polynomial way with an increase of k. With a view to establish values of D and k such that to obtain the right projections, two ways of solutions are proposed: a) We can choose any value of the determinant $D \in [10^{-50}, 10^{-10}]$ and plot the graph of Spearman's rho, by selecting the values of the parameter k from quite a wide interval. Afterwards, we should find the values of the parameter k that corresponded to the best values of

Spearman’s rho; b) In order to waste less time for visualizing data, it is recommended to plot the graphs of Spearman’s rho with several values of D, taking a shorter interval of the parameter k, and to find the values of k and D that correspond to the best values of Spearman’s rho. We should take into account that the number of neighbours must be as small as possible. The analysis of data points on three nonlinear manifolds (S-manifold, a hemisphere, “Twin peaks“) resulted in a conclusion that good projections are obtained with the small number of neighbours, as the control parameter $D = 10^{-30}$.

Analogous investigations of these datasets have also been performed using regularization algorithm R1. In all the three cases examined, the best values of Spearman’s rho (defined after investigating R2) are obtained with $t = \{10^{-3}, 10^{-2}\}$.

In this section, the best results, obtained using regularization algorithms R1 and R2 in LLE are compared. Figure 8 illustrates that similar results are obtained with these regularization algorithms in terms of Spearman’s rho. Hence both algorithms can be alternatively used in the regularization of a local Gram matrix in LLE.

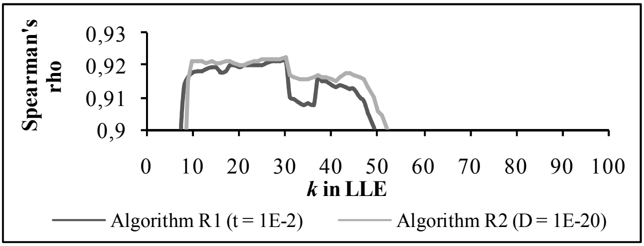


Fig. 8. Dependences of Spearman’s rho on k, obtained using regularization algorithms R1 and R2 in LLE (data points on the S-manifold were analysed)

6. Investigation of Realizations of the Laplacian Eigenmaps Method

In Section 6, the Laplacian Eigenmaps method (LE), one more nonlinear dimensionality reduction and manifold learning technique, is explored in detail and modified.

The LE algorithm, proposed by Belkin and Niyogi (2002), has two control parameters: the number k of the nearest neighbours of each data point and the heat kernel parameter T that is used in the Gaussian kernel function for computing the weights. In literature it is stated that the principle of choosing the parameter T is unknown. Belkin and Niyogi (2003) have shown in the experiments with several real datasets that it is possible to choose $T = \infty$. The algorithm has only one parameter k in this case and it works well in practice. With a view to make sure whether this statement is right in the general case, the investigation with 2000 three-dimensional data points, sampled from a nonlinear S-manifold, has been performed in this section. It has been noticed that if $T = \infty$, then the S-manifold is unfolded only as $k \leq 50$. As $k \geq 100$, we fail to unfold it. Thus, we cannot always choose $T = \infty$. It is still necessary to take count of the number of the nearest neighbours k. We see from the experiments with the points of the S-manifold that the quality of the projections obtained greatly depends on the chosen values of the parameters k and T. Spearman's rho is used to evaluate the quality of the projections. It has been defined that, in order to unfold the manifold as well as possible by increasing the number of the nearest neighbours k, the value of T must be decreased. Thus, it is difficult to choose T correctly for very large values of k. As a result, we aimed to modify the LE algorithm so that there were no need to choose the value of T, but it would be calculated while running the algorithm.

In this Section, we have proposed a modification of the LE algorithm that differs from the original LE algorithm by parameters that influence the mapping quality. Our modified LE algorithm has the following parameters:

a) The parameter w^*, $w^* \in R, 0 < w^* < 1$. It is a threshold over which the weights are considered proper, i.e., the points with such weights have influence on the considered point and can be regarded as neighbours. This parameter is a constant, because its value does not change in the run of the algorithm.

b) The maximal number of the nearest neighbours k^*. The number of the nearest neighbours for each data point is variable. However, a possibility to limit the number of neighbours is provided, taking into account the logical condition that each data point can have no more than k^* nearest neighbours. The value of k^* is chosen and it does not change in the run of the algorithm.

c) The heat kernel parameter T ($T \in R$). Any real number may be the initial value of the heat kernel parameter T, because it acquires the optimal value while running the algorithm. Thus, there is no need to choose the exact value of T, because it is set automatically, such that the number of neighbours of any point X_i would not exceed k^*.

The main idea of the LE modification is as follows. At first, we calculate the weights among all the data points. Then, based on the obtained weight matrix and the chosen threshold, we find the neighbours of each data point and construct a weighted graph.

The investigation has been performed seeking to find out how the quality of the projections obtained depends on different values of the parameter w^*. The result was that with a properly chosen number k^* of neighbours, any real number from the interval $(0, 1)$ can be the value of w^*. However, in order to achieve the highest possible quality of visualization, it is reasonable to choose the value of w^* as low as possible, for example, 0.1.

7. Topology Preservation Measures in the Visualization of Manifold-type Multidimensional Data

An important point of a manifold is its topology. There are a lot of different measures for topology preservation in the literature. An important issue solved in this thesis is to find and investigate measures such that would be suitable to analyse the topology preservation of a manifold after its embedding into a low-dimensional space. In this section, we have investigated and compared three topology preservation measures: Spearman's rho, Konig's measure (KM), and mean relative rank errors (MRRE). We have also highlighted advantages of KM and MRRE in comparison with Spearman's rho.

It is shown in Section 5 that Spearman's rho is suitable to estimate the topology preservation after transferring data points that lie on the two-dimensional manifolds to the two-dimensional space by the LLE algorithm. When calculating the values of pairwise distances between the original data points, it is necessary to use geodesic distances with a selected rather small number of neighbours (≤ 10).

The topology preservation measure KM has two control parameters: a smaller number k_1 of the nearest neighbours and a larger number k_2 ($k_1 < k_2$) of the nearest neighbours of each data point. With a view to analyse the influence of these parameters on the obtained value of KM, several investigations have been pursued in this section. KM dependences on the LLE parameter k have been calculated under different combinations of k_1 and k_2. We have established that the parameters k_1 ir k_2 influence only the magnitude of the KM value, while the form of KM dependence on k remains approximately the same. Therefore, any of these dependences can be used while looking for a number k (or its interval) of the nearest neighbours in the LLE algorithm such that a manifold be successfully unfolded in a low-dimensional space.

We have also made the analysis of the parameter K of mean relative rank errors. K denotes the number of the nearest neighbours of each data point. The results have shown that the value of this parameter may be any natural number in the interval $[5,100]$.

Two criteria have been used in the comparative analysis of the three topology preservation measures – the topology preservation quality and computational expenditure.

After investigating the data points, sampled from the manifolds of a simpler structure, such as an S-manifold, a hemisphere, a manifold “Twin peaks”, and real pictures, we have noticed that all the three measures – Spearman’s rho, KM, and MRRE – can be successfully applied to estimate the topology preservation of these manifolds, after visualizing their data points on the plane by LLE. Figure 9 demonstrates the dependences of these measures on the LLE parameter k, obtained after visualizing the pictures of a rotating duckling by LLE (Fig. 10). As far as the object has been rotated gradually at the 360° angle, the true projections have been obtained in case a) with $k \in [2,8]$ (the two-dimensional points are arranged in a circle). It is obvious that all the three dependences acquire their best values in this interval. The range of Spearman’s rho is $[-1,1]$, and the range of KM and MRRE is $[0,1]$. The best value of Spearman’s rho and KM is equal to 1, and for MRRE it is equal to 0. However, after investigating the data points sampled from the manifolds of more complex structures, such as “Swiss roll”, “Punctured sphere”, we have defined that only KM and MRRE are suitable for estimating the topology preservation of these manifolds.

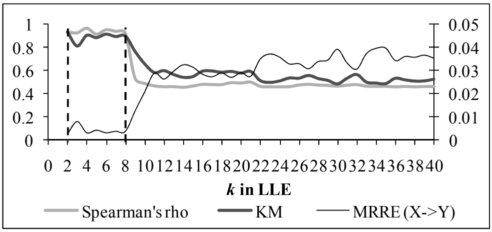


Fig. 9. Dependences of Spearman’s rho, KM, and MRRE on the LLE parameter k, obtained after visualizing the pictures of a rotating duckling on the plane by LLE

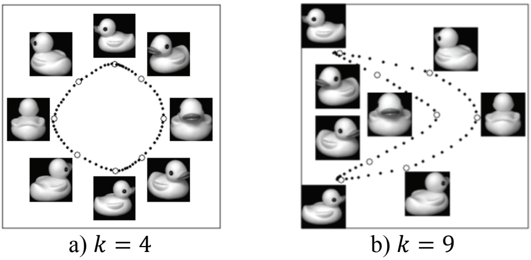

a) $k = 4$ b) $k = 9$

Fig. 10. 2-dimensional projections of 72 pictures of a rotating duckling, obtained by LLE using k nearest neighbours. The bigger circles represent pictures beside them

In order to compare the above-mentioned topology preservation measures in terms of time, the following investigation was performed: data points on various manifolds of different structure and density were analysed, their projections on a plane were obtained by the LLE algorithm, and time necessary to calculate the measures – Spearman's rho, KM, and MRRE – was determined. It has been established that the KM and MRRE measures are calculated much faster than Spearman's rho. The shorter calculation time is a great advantage of KM and MRRE as compared with Spearman's rho.

8. General Conclusions

1. The investigation of the triangulation method and Sammon‘s algorithm as well as their combination has revealed the following results:

- It does not suffice to use the triangulation method alone to visualize multidimensional data, because, after investigating experimentally the realizations of the triangulation method, where the second nearest neighbour approach and the reference point approach were used to select the reference points, it has been established that, in both cases, the projection error strongly depends on the order of points to be mapped. Besides, the projection error is rather great.
- The triangulation method is fast enough, but the projection error is great. The projection error obtained by Sammon's algorithm is not so great, however the algorithm is rather slow. It is worth using a combination of Sammon and triangulation methods, if it is necessary to map new data points of the dataset analysed quickly without losing much accuracy.

2. After investigating the relational perspective map (RPM), the following conclusions have been drawn:

- The results of the original RPM algorithm strongly depend on the parameters w (the width of the rectangle), h (the height of the rectangle), and $\tilde{r}$ (the initial learning speed). Unfortunately, there are no specific rules to select the values of these parameters.
- In our RPM modification, we avoid a strong dependence on the parameters w and h. The experiments have shown that, in the case of our algorithm, the relative value of the potential energy changes by 0.4 % at the maximum, while in the case of the original RPM algorithm, it may change up to 27%. However, the potential energy does not converge to the minimum after eliminating the parameter $\tilde{r}$. When pursuing the research, we have defined that the optimization process stabilizes after 100 iterations.

3. The investigation of locally linear embedding (LLE) has shown that:

- The parameters of the LLE algorithm – the number of the nearest neighbours of each data point and the regularization parameter of a local Gram matrix – greatly influence the quality of the visualized data.
- The new way proposed for selecting the number of the nearest neighbours allowed us to estimate a proper interval of the number of the nearest neighbours. The experiments have corroborated that the quantitative measure – Spearman's rho – is suitable to estimate the topology preservation after transferring the data to the two-dimensinal space by the LLE algorithm. In order that Spearman's rho properly reflected the projections obtained, when calculating its value it is necessary to evaluate the geodesic but not Euclidean distances in an n-dimensional space by selecting rather a small number of neighbours in the geodesic distance calculation algorithm (≤ 10).
- The proposed regularization algorithm (R2) of a local Gram matrix yields similar results as the algorithm (R1), proposed by Roweis and Saul. Hence both algorithms can be alternatively used in the regularization of the local Gram matrix in LLE. Both algorithms provide similar opportunities for analysis, each of them has one control parameter: t, in the case of R1, and D, in the case of R2. However, the advantage of D over t is that D has a real sense (it is a determinant of the regularized Gram matrix), and t is only a certain multiplier.

4. After investigating the Laplacian Eigenmaps (LE) method in detail, we can conclude that:

- The quality of the projections, obtained by the LE algorithm, greatly depends on the chosen values of the control parameters – the number k of the nearest neighbours of each data point and the heat kernel parameter T that is used in the Gaussian kernel function for computing the weights. It is difficult to choose T correctly for very large values of k.
- The proposed modification of the LE algorithm has three control parameters (T, k^*, w^*). The advantage of this modification is that there is no need to choose the value of the parameter T, because it is possible to evaluate it automatically. When a proper maximal number k^* of neighbours is chosen, any real number from the interval $(0, 1)$ can be the value of w^*. However, in order to achieve the highest possible quality of visualization, it is reasonable to choose the value of w^* as low as possible, for example, 0.1. Thus, only one significant control parameter selected – the maximal number of the nearest neighbours k^* – remains in our

modification. The variable (not fixed in advance) number of the nearest neighbours for each data point is a uniqueness of this modification.

5. In order to quantitatively estimate the topology preservation of a manifold after unfolding it in a low-dimensional space, a quantitative numerical measure must be used. After investigating three topology preservation measures – Spearman's rho, Konig's measure (KM), and mean relative rank errors (MRRE) – it was established:

- All the three measures – Spearman's rho, KM, and MRRE – can be successfully applied to estimate the topology preservation of two-dimensional manifolds of a simpler structure, after transferring their points to a plane by LLE. However, after investigating the points of the manifolds of more complex structures, we have discovered that only KM and MRRE are fit for estimating the topology preservation of these manifolds.
- Calculation of KM and MRRE is considerably faster as compared with Spearman's rho because these criteria use the Euclidean distances only, while Spearman's rho uses the geodesic distances that are more computationally expensive. Moreover, KM and MRRE evaluate a limited number of neighbours of each point from the dataset analysed, while Spearman's rho considers distances between all the pairs of points from the dataset analysed. Therefore, it tries to take into account the global structure of the manifold. However, in some cases, it may be not optimal, because some local properties of the manifold may be lost.

The List of Author‘s Scientific Publications on the Subject of the Dissertation

Articles in scientific journals reviewed:

1. Karbauskaitė, R.; Dzemyda, G. 2009. Topology preservation measures in the visualization of manifold-type multidimensional data. *Informatica*, 20(2), 235–254. ISSN 0868-4952 (*ISI Web of Science*).
2. Karbauskaitė, R.; Kurasova, O.; Dzemyda, G. 2007. Selection of the number of neighbours of each data point for the locally linear embedding algorithm. *Information Technology and Control*, 36(4), 359–364. ISSN 1392-124X (*ISI Web of Science*).
3. Karbauskaitė, R.; Dzemyda, G. 2006. Multidimensional data projection algorithms saving calculations of distances. *Information Technology and Control*, 35(1), 57–64. ISSN 1392-124X (*VINITI, INSPEC*).
4. Karbauskaitė, R.; Marcinkevičius, V.; Dzemyda, G. 2006. Testing the relational perspective map for visualization of multidimensional data.

Technological and Economic Development of Economy, 12(4), 289–294. ISSN 1392-8619 (*ASCE Civil Engineering Abstracts, Business Source Complete, Business Source Premier, Current Abstracts, ICONDA, SCOPUS, TOC Premier*).

Articles in the proceedings of scientific conferences:

5. Karbauskaitė, R.; Dzemyda, G. 2009. Dependence of the Laplacian Eigenmaps method and its modification on the parameters. *Proceedings of the 13th International Conference "Applied Stochastic Models and Data Analysis" (ASMDA-2009): selected papers.* Vilnius: Technika. 263–268. ISBN 978-9955-28-463-5.
6. Karbauskaitė, R.; Dzemyda, G.; Marcinkevičius, V. 2008. Selecting a regularization parameter in the locally linear embedding algorithm. *Proceedings of the 20th international EURO mini conference "Continuous optimization and knowledge-based technologies" (EurOPT'2008): selected papers.* Vilnius: Technika. 59–64. ISBN 978-9955-28-283-9 (*Conference Proceedings Citation Index*).

Short Description about the Author

Rasa Karbauskaitė was born in Veiviržėnai on the 29th of December in 1980. After finishing the Gargždai "Vaivorykštė" gymnasium in 1999, she graduated from the Vilnius Pedagogical University in 2003 acquiring a bachelor's degree in mathematics, and in 2005, she acquired a master's degree in informatics with a magna cum laude diploma. Since 2006 she has been a PhD student at the Institute of Mathematics and Informatics.

E-mail: karbauskaite@ktl.mii.lt

DAUGIAMAČIŲ DUOMENŲ VIZUALIZAVIMO METODŲ, IŠLAIKANČIŲ LOKALIĄ STRUKTŪRĄ, ANALIZĖ

Problemos aktualumas

Nėra tokios žmonių veiklos srities, kur nebūtų kaupiami ir analizuojami daugiamačiai duomenys. Tipiškas daugiamačių duomenų pavyzdys susijęs su vaizdų apdorojimu. Dažnai duomenis sudaro to paties objekto paveikslėliai, gauti palaipsniui pasukant objektą tam tikru kampu arba nufotografuojant jį skirtingais momentais. Kiekvienas paveikslėlis yra skaitmenizuojamas, t.y. duomenų taško koordinatės yra sudarytos iš paveikslėlio taškų spalvinių savybių, ir todėl šio taško koordinačių skaičius yra labai didelis. Duomenų dimensijos mažinimo metodai (projekcijos metodai) vystomi gana intensyviai. Transformavus jais daugiamačius duomenis į dvimatę ar trimatę vaizdo erdvę ir juos vizualizavus, daug paprasčiau suvokti duomenų struktūrą ir sąryšius tarp jų. Tačiau duomenis transformuojant į mažesnės dimensijos erdvę, neišvengiami duomenų projekcijų iškraipymai. Todėl gautų projekcijų kokybės įvertinimas išlieka aktualia problema.

Dažnai tenka dirbti su duomenų aibėmis, kurios pastoviai papildomos naujais duomenimis. Labai svarbu greitai atvaizduoti naujus duomenų taškus, neprarandant didelio tikslumo. Todėl naujų taškų atvaizdavimas, jų įterpimas tarp anksčiau atvaizduotų taškų – viena iš disertacijoje nagrinėjamų problemų.

Daugiamačių duomenų projekcijos metodai susiduria su dviem pagrindinėmis problemomis. Pirma, reikia rasti daugiamačių duomenų projekcijas mažesnės dimensijos erdvėje (dvimatėje ar trimatėje), siekiant kuo tiksliau išlaikyti analizuojamos aibės objektų artimumus – panašumus ar skirtingumus. Antra, daugiamačius duomenis atvaizduoti mažesnės dimensijos erdvėje taip, kad jų projekcijos nepersidengtų. Ši problema taip pat yra viena iš disertacijoje sprendžiamų problemų.

Dažnai praktiniuose uždaviniuose yra sukaupiami daugiamačiai duomenys, kuriuos atitinkantys taškai nagrinėjami didelės dimensijos erdvėje, o iš tikrųjų jie yra kokios nors mažesnės dimensijos daugdaros arba tai daugdarai artimi taškai. Taigi viena iš pagrindinių disertacijos problemų – atrasti mažesnio matavimo netiesinę daugdarą didelio matavimo erdvėje ir tada transformuoti duomenų taškus, esančius ant arba arti tos daugdaros, į mažesnio matavimo erdvę.

Svarbus su daugdara susijęs dalykas yra jos topologija. Topologijos išlaikymui įvertinti sukurta daugybė įvairių matų. Svarbi disertacijoje sprendžiama problema – rasti ir ištirti tuos matus, kurie būtų tinkamiausi analizuoti daugdaros topologijos išlaikymą po jos transformavimo į mažesnės dimensijos erdvę.

Tyrimo objektas

Disertacijos tyrimų objektas – daugiamačių duomenų vizualizavimo algoritmai ir metodai, išlaikantys lokalią struktūrą, bei daugiamačių duomenų projekcijų mažesnės dimensijos erdvėje vertinimo kriterijai. Čia lokalios struktūros išlaikymu vadiname atstumų tarp artimiausių taškų santykių išlaikymą po analizuojamos daugiamačių duomenų aibės transformavimo iš didesnio matavimo erdvės į mažesnio matavimo erdvę.

Darbo tikslas ir uždaviniai

Disertacijos tikslas yra išanalizuoti daugiamačių duomenų vizualizavimo algoritmus, išlaikančius lokalią struktūrą, juos modifikuoti bei ištirti nagrinėjamų algoritmų valdymo parametrų svarbą ir pasiūlyti būdus šiems parametrams parinkti, siekiant gauti tikslesnę duomenų projekciją.

Norint pasiekti šį tikslą, reikėjo išspręsti tokius uždavinius:

1. išanalizuoti esamus daugiamačių duomenų vizualizavimo metodus ir apsibrėžti tiriamų vizualizavimo metodų, išlaikančių lokalią struktūrą, grupę;
2. išanalizuoti pasirinktus metodus ir jais vizualizuoti daugiamačių duomenų aibes;
3. tiriamais metodais gautus vizualizavimo rezultatus palyginti su rezultatais, kurie gauti metodais, išlaikančiais ne tik lokalią struktūrą;
4. įvertinti įvairiais vizualizavimo metodais gautų duomenų projekcijų kokybę (naudojant ekspertinį ir kiekybinius skaitinius matus);
5. sukurti daugiamačių duomenų vizualizavimo algoritmų, išlaikančių lokalią struktūrą, modifikacijas, siekiant gauti tikslesnę analizuojamų duomenų projekciją;
6. įvertinti modifikuotų algoritmų rezultatus lyginant su rezultatais, gautais originaliais algoritmais;
7. ištirti nagrinėjamų algoritmų valdymo parametrų svarbą ir pasiūlyti būdus šiems parametrams parinkti, siekiant gauti tikslesnę analizuojamų duomenų projekciją.

Mokslinis naujumas

1. Eksperimentiškai ištyrus trianguliacijos metodo realizacijas, naudojančias antrojo arčiausiojo kaimyno ir atramos taško metodus atraminiams taškams parinkti, nustatyta, jog abiem atvejais projekcijos paklaida labai priklauso nuo taškų atvaizdavimo sekos, o tai įrodo, kad naudoti vien tik trianguliacijos metodą duomenims vizualizuoti nėra pakankama. Pasiūlyta Sammono ir trianguliacijos metodų junginio nauja realizacija naujiems taškams atvaizduoti.
2. Eksperimentiškai ištirta santykinės perspektyvos metodo (RPM) priklausomybė nuo parametrų – stačiakampio, kuriame vizualizuojami duomenys, pločio ir aukščio, bei pradinio mokymo greičio. Pasiūlyta nauja RPM metodo realizacija, leidžianti šios priklausomybės beveik išvengti.
3. Eksperimentiškai ištirta lokaliai tiesinio vaizdavimo (LLE) metodo priklausomybė nuo parametrų – artimiausių kaimynų skaičiaus kiekvienam duomenų taškui ir lokaliosios Gramo matricos reguliarizacijos parametro. Siekiant gauti kuo tikslesnes duomenų projekcijas, pasiūlytas naujas būdas artimiausių kaimynų skaičiui parinkti LLE algoritme. Taip pat sukurtas naujas algoritmas lokaliajai Gramo matricai reguliarizuoti.
4. Eksperimentiškai ištirta parametrų (artimiausių kaimynų skaičiaus kiekvienam duomenų taškui ir šiluminio branduolio parametro, naudojamo Gauso branduolio funkcijoje svoriams apskaičiuoti) svarba Laplaso matricos tikrinių žemėlapių (LE) algoritme. Pasiūlyta LE algoritmo modifikacija, kurioje yra tik vienas svarbus valdymo parametras – maksimalus artimiausių kaimynų skaičius.
5. Panaudojant du kriterijus – topologijos išlaikymo kokybę ir skaičiavimo sąnaudas – ištirti ir palyginti trys topologijos išlaikymo matai (Spirmano koeficientas, Konigo matas (KM) ir kaimynystės klaidos (MRRE)), kurie tinkami analizuoti daugdaros topologijos išlaikymą po jos transformavimo į mažesnio matavimo erdvę.

Ginamieji teiginiai

1. Taupant skaičiavimų laiką ir mažai teprarandant tikslumą, naujiems daugiamatės erdvės taškams atvaizduoti, pradinius taškus vizualizavus MDS tipo metodu, gali būti naudojamas trianguliacijos metodas.
2. Santykinės perspektyvos, lokaliai tiesinio vaizdavimo ir Laplaso matricos tikrinių žemėlapių metodų valdymo parametrai labai įtakoja gautų projekcijų kokybę, tačiau egzistuoja strategijos, leidžiančios mažinti parametrų skaičių ar reglamentuoti jų reikšmių parinkimą.

3. Konigo matas (KM) ir kaimynystės klaidos (MRRE) visada gerai nusako daugdaros topologijos išlaikymą po jos transformacijos į mažesnio matavimo erdvę, o Spirmano koeficientas sėkmingai gali būti taikomas tik paprastesnės struktūros daugdarų topologijos išlaikymui įvertinti.

Praktinė vertė

Tyrimų rezultatai atskleidė daugdaros tipo daugiamačių duomenų analizės galimybes. Parodyta, jog netiesinės daugdaros atpažinimo metodai gali būti plačiai naudojami įvairiose srityse, tarp jų ir medicinoje.

Tyrimai atlikti pagal Lietuvos valstybinio mokslo ir studijų fondo aukštųjų technologijų plėtros programos projektą „Informacinės klinikinių sprendimų palaikymo ir gyventojų sveikatinimo priemonės e. Sveikatos sistemai (Info Sveikata)“; Registracijos Nr.: B-07019; Vykdymo laikas: 2007 m. 09 mėn. – 2009 m. 12 mėn.

Darbo rezultatų aprobavimas ir publikavimas

Tyrimų rezultatai publikuoti 6 moksliniuose leidiniuose (keturi iš jų žurnaluose): 2 straipsniai tarptautiniuose periodiniuose leidiniuose, įtrauktuose į Mokslinės informacijos instituto pagrindinį sąrašą (*ISI Web of Science*), 2 straipsniai kituose recenzuojamuose mokslo žurnaluose, 2 straipsniai konferencijų pranešimų rinkiniuose.

Tyrimų rezultatai buvo pristatyti ir aptarti 5 nacionalinėse ir tarptautinėse konferencijose.

Darbo apimtis

Disertaciją sudaro 8 skyriai ir literatūros sąrašas. Bendra disertacijos apimtis 168 puslapiai, 103 paveikslai ir 6 lentelės. Pirmame skyriuje atskleidžiama nagrinėjamos problematikos svarba, įvardinamas tyrimų objektas, aprašomi keliami tikslai bei uždaviniai, mokslo naujumas ir kt. Antras skyrius skirtas daugiamačių duomenų projekcijos metodų analitinei apžvalgai. Ji leido išgryninti kelias aktualias tyrimų kryptis, kurias vienija lokalios struktūros išlaikymo būtinumas. Trečiame skyriuje lokalios struktūros išlaikymo idėjos pritaikomos duomenų aibių papildymui naujais duomenimis. Ketvirtame skyriuje nagrinėjamas metodas, leidžiantis daugiamačius duomenis atvaizduoti mažesnės dimensijos erdvėje taip, kad jų projekcijos nepersidengtų. Čia irgi akivaizdi lokalios struktūros išlaikymo panaudojimo idėja. Penktas ir šeštas skyriai skirti nagrinėti dviem netiesinės daugdaros atpažinimo metodams, kurie daugiamačius duomenis transformuoja į mažesnės dimensijos erdvę, išlaikant kaimyniškumą tarp artimiausių taškų ir atskleidžiant daugiamačių duomenų netiesinę struktūrą. Septintas skyrius skirtas daugdaros topologijos išlaikymo matams tyrinėti. Aštuntame skyriuje pateiktos bendrosios išvados.

Bendrosios išvados

1. Trianguliacijos metodo ir Sammono algoritmo bei jų jungimo tyrimas atskleidė faktus:

- Naudoti vien tik trianguliacijos metodą daugiamačiams duomenims vizualizuoti nėra pakankama, nes, eksperimentiškai ištyrus trianguliacijos metodo realizacijas, naudojančias antrojo arčiausiojo kaimyno ir atramos taško metodus atraminiams taškams parinkti, nustatyta, jog abiem atvejais projekcijos paklaida labai priklauso nuo taškų atvaizdavimo sekos. Be to, paklaida gana didelė.
- Trianguliacijos metodas yra pakankamai greitas, bet projekcijos paklaida didelė. Sammono algoritmu gauta projekcijos paklaida nedidelė, tačiau jis yra gana lėtas. Sammono ir trianguliacijos metodų junginį verta naudoti, kai reikia greitai atvaizduoti naujus analizuojamos aibės taškus neprarandant didelio tikslumo.

2. Ištyrus santykinės perspektyvos metodą (RPM), padarytos šios išvados:

- Pradinio RPM algoritmo rezultatai labai priklauso nuo parametrų w (stačiakampio plotis), h (stačiakampio aukštis) ir $\tilde{r}$ (pradinis mokymo greitis). Deja, literatūroje nėra suformuluotų aiškių taisyklių, kaip šiuos parametrus parinkti.
- Darbe pasiūlytame algoritme nėra didelės priklausomybės nuo parametrų w ir h. Tyrimai parodė, jog naudojant mūsų algoritmą, energijos santykinė reikšmė daugiausia pakinta 0,4%, tuo tarpu, naudojant RPM pradinį algoritmą, ji gali kisti net iki 27%. Tačiau, atsisakius parametro $\tilde{r}$, potencinė energija nekonverguoja į minimumą. Tyrimų metu pastebėta, kad, atlikus apie 100 iteracijų, procesas stabilizuojasi.

3. Lokaliai tiesino vaizdavimo (LLE) metodo tyrimas parodė:

- LLE algoritmo parametrai – artimiausių kaimynų skaičius kiekvienam duomenų taškui ir lokaliosios Gramo matricos reguliarizacijos parametras – labai įtakoja duomenų vizualizavimo kokybę.
- Pasiūlytas naujas būdas artimiausių kaimynų skaičiui parinkti leido nustatyti tinkamą artimiausių kaimynų skaičiaus intervalą, o ne tik vieną skaičių. Eksperimentai parodė, kad kiekybinis matas – Spirmano koeficientas – yra tinkamas duomenų topologijos išlaikymui įvertinti,

duomenis transformavus LLE metodu į mažesnės dimensijos erdvę. Tam, kad Spirmano koeficientas tinkamai atspindėtų gautas projekcijas, skaičiuojant jo reikšmę reikia n-matėje erdvėje vertinti geodezinius, o ne Euklido atstumus, ir geodezinių atstumų skaičiavimo algoritme fiksuoti gana mažą artimiausių kaimynų skaičių (≤ 10).

- Pasiūlytas naujas algoritmas lokaliajai Gramo matricai reguliarizuoti (R2) pateikia panašius rezultatus kaip ir algoritmas (R1), pasiūlytas Roweis ir Saul. Taigi abu algoritmai gali būti alternatyviai naudojami LLE metode reguliarizuojant Gramo matricą. Abu algoritmai pateikia panašią galimybę analizei, kiekvienas iš jų turi po vieną valdymo parametrą: t R1 atveju ir D R2 atveju. Tačiau parametro D privalumas lyginant su t yra tai, kad parametras D turi realią prasmę (jis yra reguliarizuotos Gramo matricos determinantas), o t yra tiktai tam tikras daugiklis.

4. Detaliai ištyrus Laplaso matricos tikrinių žemėlapių metodą (LE), padarytos šios išvados:

- LE algoritmu gautų projekcijų kokybė labai priklauso nuo parinktų valdymo parametrų reikšmių – artimiausių kaimynų skaičiaus k kiekvienam duomenų taškui ir šiluminio branduolio parametro T, naudojamo Gauso branduolio funkcijoje svoriams apskaičiuoti. Sunku teisingai parinkti T, jei parametro k reikšmės didelės.
- Pasiūlyta LE algoritmo modifikacija, turinti tris valdymo parametrus (T, k^*, w^*). Šios modifikacijos privalumas yra tai, kad šiluminio branduolio parametro T reikšmės parinkinėti nereikia, nes sudaryta galimybė ją įvertinti automatiškai. Kai pasirinktas tinkamas maksimalus kaimynų skaičius k^*, parametro w^* reikšme gali būti bet koks realus skaičius iš intervalo $(0; 1)$. Tačiau, norint pasiekti kuo geresnę vizualizavimo kokybę, patartina pasirinkti parametro w^* reikšmę kuo mažesnę, pvz., 0,1. Taigi mūsų LE modifikacijoje lieka tik vienas svarbus valdymo parametras k^* – maksimalus artimiausių kaimynų skaičius. Kintamas (iš anksto nefiksuotas) artimiausių kaimynų skaičius kiekvienam taškui yra šios modifikacijos unikalumas.

5. Siekiant kiekybiškai įvertinti daugdaros topologijos išlaikymą, transformavus ją į mažesnės dimensijos erdvę, reikia naudoti kiekybinius skaitinius matus. Išanalizavus tris topologijos išlaikymo matus: Spirmano koeficientą, Konigo matą (KM) ir kaimynystės klaidas (MRRE), nustatyta:

- Visi trys matai – Spirmano koeficientas, KM ir MRRE – sėkmingai gali būti taikomi paprastos struktūros dvimačių daugdarų topologijos išlaikymui įvertinti, transformavus jų taškus į plokštumą LLE metodu. Tačiau, atlikus tyrimus su sudėtingesnės struktūros daugdarų taškais, paaiškėjo, jog tik KM ir MRRE tinka šių daugdarų topologijos išlaikymui įvertinti.

- KM ir MRRE matų apskaičiavimas yra žymiai greitesnis nei Spirmano koeficiento, nes šie kriterijai naudoja tiktai Euklido atstumus. Tuo tarpu, Spirmano koeficientas naudoja geodezinius atstumus, kurių apskaičiavimas reikalauja daugiau laiko sąnaudų. Be to, Spirmano koeficientas vertina visus atstumus tarp tiriamos duomenų aibės taškų, o KM ir MRRE – tarp nedidelio kaimyninių taškų skaičiaus. Taigi Spirmano koeficientas stengiasi atsižvelgti į globalią daugdaros struktūrą. Tačiau kai kuriais atvejais tai nėra optimalu, nes gali būti prarastos kai kurios daugdaros lokalios savybės.

Trumpos žinios apie autorę

Rasa Karbauskaitė gimė 1980 m. gruodžio 29 d. Veiviržėnuose. 1999m. baigė Gargždų „Vaivorykštės“ gimnaziją. 2003 m. įgijo matematikos bakalauro laipsnį ir mokytojo kvalifikaciją Vilniaus pedagoginio universiteto Matematikos ir informatikos fakultete. 2005 m. įgijo informatikos magistro laipsnį Vilniaus pedagoginiame universitete. 2006-2010 m. Matematikos ir informatikos instituto doktorantė.

El. pašto adresas: karbauskaite@ktl.mii.lt

Rasa Karbauskaitė

ANALYSIS OF MULTIDIMENSIONAL DATA VISUALIZATION METHODS THAT PRESERVE THE LOCAL STRUCTURE

Summary of Doctoral Dissertation
Physical Sciences (P 000)
Informatics (09 P)
Informatics, Systems Theory (P 175)

Rasa Karbauskaitė

DAUGIAMAČIŲ DUOMENŲ VIZUALIZAVIMO METODŲ, IŠLAIKANČIŲ LOKALIĄ STRUKTŪRĄ, ANALIZĖ

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

2010 08 10. 2 sp. l. Tiražas 60 egz.
Parengė spaudai ir išleido Matematikos ir informatikos institutas
Akademijos g. 4, LT-08663 Vilnius.
Interneto svetainė: http://www.mii.lt.

Spausdino UAB „Kauno technologijos universiteto spaustuvė“
Studentų g.54, LT-51424 Kaunas