ISSN 1392-8619 print/ISSN 1822-3613 online
ŪKIO TECHNOLOGINIS IR EKONOMINIS VYSTYMAS
TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY
http://www.tede.vgtu.lt
2006, Vol XII, No 4, 283–288

STRATEGIES OF SELECTING THE BASIS VECTOR SET IN THE RELATIVE MDS

Jolita Bernatavičienė[1], Gintautas Dzemyda[1,2], Olga Kurasova[1,2], Virginijus Marcinkevičius[1]

[1]*Institute of Mathematics and Informatics, Akademijos g. 4, LT-08663 Vilnius, Lithuania*
[2]*Vilnius Pedagogical University, Studentų g. 39, LT-08106 Vilnius, Lithuania*
E-mails: JolitaB@ktl.mii.lt, Dzemyda@ktl.mii.lt, Kurasova@ktl.mii.lt, VirgisM@ktl.mii.lt

Received 15 June 2006; accepted 20 November 2006

Abstract. In this paper, a method of large multidimensional data visualization that associates the multidimensional scaling (MDS) with clustering is modified and investigated. In the original algorithm, the visualization process is divided into three steps: the basis vector set is constructed using the *k*-means clustering method; this set is projected onto the plane using the MDS algorithm; the remaining data set is visualized using the relative MDS algorithm. We propose a modification which differs from the original algorithm in the strategy of selecting the basis vectors. In our modification, the set of basis vectors consists of vectors that are selected from *k* clusters in a new way. The experimental investigation showed that the modification exceeds the original algorithm in visualization quality and computational expenses.

Keywords: multidimensional scaling, visualization, clustering, basis vector set, new points mapping.

1. Introduction

Objects from the real world are often described by some attributes (parameters). If these attributes are numerical ones, it is possible to form multidimensional vectors, corresponding to each analysed object. Denote the multidimensional vectors by $X^1, X^2, ..., X^m$ ($X^i = (x_1^i, x_2^i, ..., x_n^i)$, $i = 1, ..., m$). Here m is the number of the analysed objects, n is the number of attributes of the objects. A human being can comprehend visual information easier and more quickly than the numerical one. So, it is useful to present multidimensional vectors in some visual form. Various methods can be used for this purpose. It is possible to divide them into two groups: (1) dimension reduction methods (principal component analysis [1], projection pursuit [2], multidimensional scaling [3], etc.); and (2) methods based on neural networks (self-organizing maps (SOM) [4], combination of the SOM and Sammon's mapping [5], etc.).

The multidimensional scaling method (MDS) [3] is a popular method usable to visualize multidimensional data. However, we face some problems when we have to project (visualize) a large data set or to map a new data point among the previously mapped points. In the MDS method, every iteration requires each point to be compared with all the other points and the iteration complexity is $O(m^2)$. Thus, the MDS method is unsuitable for large data sets: it takes much computing time or there is not enough computing memory. Furthermore, it is necessary to recalculate the projection of all data points, when a point has to be mapped. Various methods have been suggested for mapping of new points without recalculating all the previously mapped points: Sammon's mapping based on an artificial neural network (SAMANN) [6], simple two-dimensional mapping [7], distance mapping [8], incremental scaling [9], relative MDS [10], and neuroscale [11].

In this paper, we focus on the relative MDS method and analyse strategies of the selecting the basis vector set. One strategy has been proposed and analysed in [12], that is based on the results of *k*-means algorithm. We propose two other superior strategies.

2. Data analysis methods

The 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.

$$E_{MDS} = \sum_{\substack{i,j=1 \\ i<j}}^{m} (d_{ij}^* - d_{ij})^2. \qquad (1)$$

Various types of minimization of the stress function are possible [3], [13]. In this paper, we use the SMACOF algorithm based on iterative majorization. It is one of the best optimisation algorithms for this type of minimization problem [14]. This method is simple and powerful, because it guarantees a monotone convergence of stress function [3], [14].

Relative MDS is proposed in [10]. In classification tasks, it may be interesting to see where a new data point "falls" among the known cases and discover the class topology of its neighbouring known cases to get an insight on how a classifier would classify this new point. The realization of this purpose gives rise to the need for a method that allows the mapping of one new point on a set of data points previously mapped, using the topology-preserving mapping. The MDS is a topology preserving mapping, but it does not offer a possibility 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* in 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 the moving points, and interpoint distances between the moving points. The stress function (1) is rewritten as:

$$E_{Relative\_MDS} = \sum_{\substack{i,j=1 \\ i<j}}^{N_{new}} (d_{ij}^* - d_{ij})^2 + \sum_{i=1}^{N_{new}} \sum_{j=N_{new}+1}^{N_{total}} (d_{ij}^* - d_{ij})^2. \qquad (2)$$

In our experiments, we use the Quasi-Newton algorithm to minimize $E_{Relative\_MDS}$.

The *k*-means method is an iterative clustering algorithm in which the analysed vectors are moved among the sets of clusters until the desired set is reached [15]. Let the set of vectors mapped to the ith cluster be $\{X^{i1}, X^{i2},..., X^{i\mu_i}\}$. Here μ_i is the number of the objects in the *i*th cluster ($X^{ij} = (x_1^{ij}, x_2^{ij},..., x_n^{ij})$, $j = 1,..., \mu_i$). The squared error is defined as:

$$E_k = \sum_{i=1}^{k} \sum_{j=1}^{\mu_i} \| X^{ij} - C_i \|^2. \qquad (3)$$

Here $C_i = (c_1^i, c_2^i,..., c_n^i)$ is the centre of the cluster, ($c_k^i = \frac{1}{\mu_i} \sum_{j=1}^{\mu_i} x_k^{ij}$, $k = 1,..., n$).

3. Data sets for analysis

In the experiments, the data set, obtained using the ellipsoidal cluster generator [16], is used. This generator creates ellipsoidal clusters with the major axis of an arbitrary orientation. The boundary of a cluster is defined by four parameters:

1. the origin (which is also the first focus),
2. the interfocal distance, uniformly distributed in the range [1.0, 3.0],
3. the orientation of the major axis, uniformly located amongst all orientations,
4. the maximum sum of Euclidean distances to two foci, depending to the range [1.05, 1.15] – equivalent to the eccentricity ranging from [0.870, 0.952].

For each cluster, data points are generated at a Gaussian-distributed distance from a uniformly random point on the major axis, in a uniformly random direction, and are rejected if they lie outside the boundary. Using this ellipsoidal generator, 1115 50-dimensional points are generated that form 20 clusters. This set is used in the experiments of comparative analysis.

The other data set consists of the points of two hyper-

spheres. 100 10-dimensional points are in each sphere. This set is used to illustrate strategies of selection of the basis vector set.

4. Strategies of selection of the basis vector set

In the relative MDS, there arises a problem of selection of the basis vector set *F*. Some strategies can be used:

I. Set *F* consists of the cluster centres, obtained by *k*-means clustering algorithm.

II. Set *F* consists of data set points that are the closest points to the cluster centres, obtained by *k*-means algorithm. Additional points of each cluster are added to set *F*: these points are selected to be farthest from the respective cluster centres.

III. Set *F* consists of data set points, chosen randomly from the whole data set.

Strategy I was proposed and analysed in [12]. We propose here Strategies II and III. The general scheme of visualization process is presented in Fig 1.

In Fig 2, the projections of two spheres are pre-sented for illustration of the strategies. At first, it is necessary to form the basis vector set *F*. The points of the formed basis set *F* are mapped on the plane, using the MDS algorithm. Using Strategy I, set *F* consists of the centres of 20 clusters (marked by circled crosses in Fig 2 a). Using Strategy II, set *F* consists of two sub-sets: (a) two points (marked by circled crosses in Fig 2 b), which are closest to the centres of two clusters and (b) 9 points of each cluster (marked by unfilled circles in Fig 2 b). Using Strategy III, set *F* consists of 20 points from the data set chosen randomly: the visualization results are very similar to these of Fig 2 a.

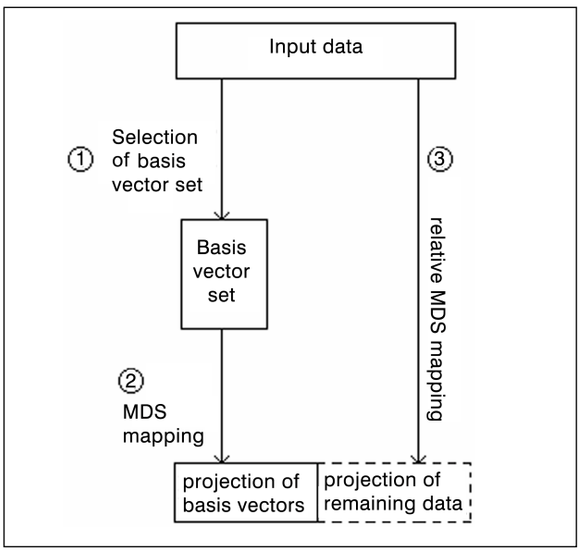


Fig 1. Scheme of the visualization process: (1) selection of the basis vector set, (2) the basis vector set is projected by MDS mapping, (3) the remaining points are projected by relative MDS mapping

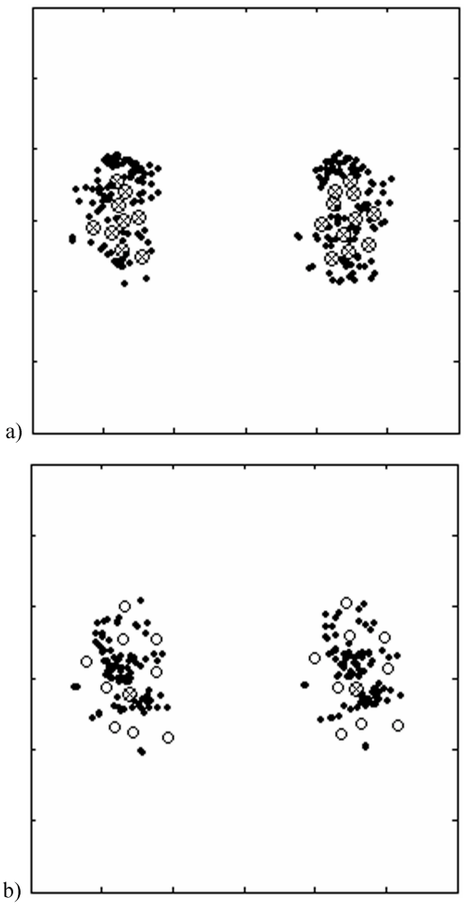


Fig 2. Projections of two spheres: (a) using Strategy I $E = 0.12814$, (b) using Strategy II $E = 0.12265$

The number of the basis vectors N_{fixed} is equal to 20 in all the three cases (Strategies I, II, and III). Then the remaining points (set *M*), marked by filled circles, are mapped by the relative MDS algorithm.

To compare the obtained visualization results, the projection error is calculated:

$$E = \sqrt{\sum_{i<j}^{m} (d_{ij}^{*} - d_{ij})^2 \Big/ \sum_{i<j}^{m} (d_{ij}^{*})^2}. \qquad (4)$$

The projection error *E* (4) is used here instead of E_{MDS} (1), because the inclusion of the normalized parameter $\sum_{i<j}^{m} (d_{ij}^{*})^2$ gives a clear interpretation of image quality that does not depend on the scale and the number of distances in an *n*-dimensional space. The reason for using *E* rather than the squared error E^2 is that E^2 is almost always very small in practice, so *E* values are easier to discriminate [3]. Of course, the error *E* (4) may be used in the MDS, however, it is impossible to decompose and apply this error for

the relative MDS. Therefore, E_{MDS} (1) is minimized.

Using Strategy II, the projection error (4) is obtained a little smaller ($E = 0.12265$) than using Strategy I ($E = 0.12814$) (Fig 2).

5. Results of comparative analysis

With a view to evaluate the influence of a basis vector set on the visualization results, three strategies, mentioned above, are compared.

Using Strategy I, the basis vector set is constructed from centres of clusters. An experiment consists of two stages of the whole data set visualization: (a) the mapping of the basis vector set using the MDS algorithm and (b) the mapping of the remaining data using the relative MDS algorithm. As the results of the *k*-means clustering algorithm depend on the selection of initial centres of the clusters, the experiment has been repeated for 10 times choosing different sets of initial cluster centres for each fixed number of clusters $N_{fixed} = 100, 200, ..., 800$.

Using Strategy II, the experiments are performed with the following number of clusters: 10, 20,…, 80. Here the nearest points to the cluster centres plus 9 or less additional points from each cluster are used to form the basis vector set *F*. The number of additional points depends on clustering results: one or some clusters, obtained by the *k*-means algorithm, may consist of less than 9 points. So, the number of the basis vectors N_{fixed} is almost equal to 100, 200,…, 800. Each experiment has been repeated for 10 times choosing different sets of the initial cluster centres.

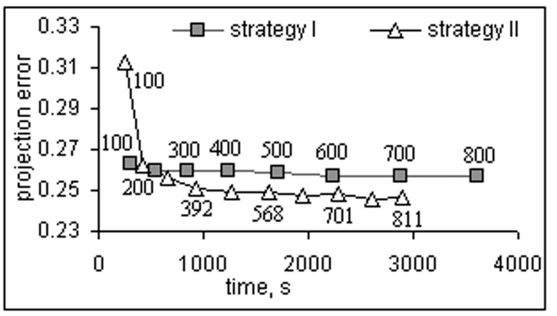
Using Strategy III, the experiments are done with the following number of the data set points chosen randomly: $N_{fixed} = 100, 200, ..., 800$. Each experiment has been repeated for 10 times choosing a different set of points.

The projection error (4) is used to estimate the visualization results. These errors, obtained in 10 experiments, are averaged for each strategy individually. The dependence of the projection error on computing time is presented in Fig 3. Numbers, marked near the curves, denote the numbers of the basis vectors N_{fixed}. Fig 3 a shows that the lower projection error is obtained by using Strategy II than using Strategy I. The computing time is saved: a lower projection error is obtained faster even with larger number of the basis vectors. When all points ($m = 1115$) of the data set are mapped by MDS, the computing time is 13 511 s, and $E = 0.26554$ (after 200 iterations). Using the modification of the relative MDS, the projection error is lower, and the computing time is saved significantly (4000 s, $E = 0.25670$ in the worst case). Fig 3 b shows that the projection errors are very similar, using Strategies II and III. This fact proves the efficiency of the relative MDS for mapping the large data set.

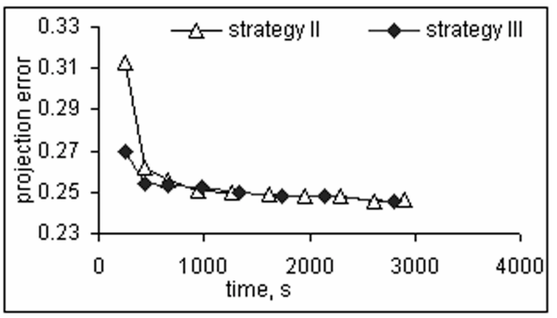
The dependence of the projection error on the number of the basis vectors N_{fixed} is presented in Fig 4. It shows that the averaged projection error *E* constantly decreases, when N_{fixed} increases, using Strategies II and III. Using Strategy I, the averaged projection error *E* stabilizes for $N_{fixed} \approx 300$, and with an N_{fixed} increase, *E* changes inessentially. The error *E*, obtained using Strategy I, is greater than that, obtained using Strategies II and III.

The projection errors, obtained by using a different number of N_{fixed} (the number of clusters $k = 10, 20, ..., 50$, and the number of the data set points from each cluster $p = 5, 10, ..., 25$) are presented in Table 1 and Fig 5 for Strategy II.

The experiments illustrate that the number *k* of clusters has influence on the projection error and on the visualization results (Fig 6). In this case (Table 1, Fig 2), it is necessary to use more than 30 clusters. When the number *p* of



a)



b)

Fig 3. Dependence of the projection error on computing time

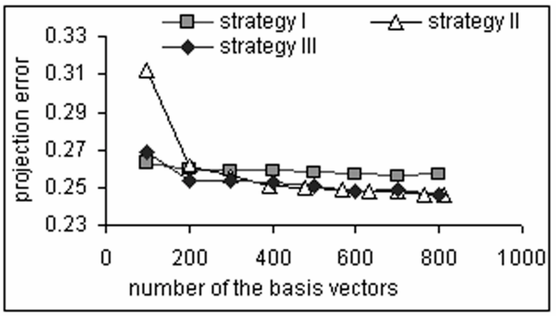


Fig 4. Dependence of the projection error on the number of the basis vectors

the data set points from each cluster increases, the projection error decreases and the visualization quality improves. However, too large number N_{fixed} of the basis vectors increases the computing time, while the error changes inessentially. In Fig 6, the visualization results of the ellipsoidal data are presented: (a) $k = 10$, $p = 10$, (b) $k = 50$, $p = 10$. The lower projection is obtained and the quality of visualization is better in case (b).

6. Conclusions

The visual analysis of large data sets is a topical problem. However, when a large data set of multidimensional vectors is visualized by the MDS method, it takes much computing time. In this paper, we have investigated a modification of the MDS method for large data sets: at first, some basis vectors are projected onto the plane, then the remaining points are projected among the previously mapped points.

The investigation allows to draw the following conclusions:

- the strategy of selecting the basis vectors directly influences the visualization results;
- the better visualization results are obtained when the basis vectors are selected so that they cover the area of location of analysed data set as uniformly as possible;
- 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;
- it is expedient to take the basis vectors from the data points which are closest to the cluster centres instead of direct selection of cluster centres as the basis vectors.

The efficiency of Strategy II and III is similar for the ellipsoidal data set. Further investigations should be pursued with the larger number of data sets in order to get the exact estimates on comparative efficiency of these two strategies.

References

1. Taylor, P. Statistical Methods. In: Intelligent Data Analysis: an Introduction, edited by Berthold M., Hand D. J., Springer-Verlag, 2003, p. 69–129.
2. Brunsdon, C.; Fotheringham, A. S.; Charlton, M. E. An Investigation of Methods for Visualising Highly Multivariate Datasets. In: Case Studies of Visualization in the Social Sciences, edited by Unwin D., Fisher P., 1998, p. 55–80.
3. Borg, I.; Groenen, P. Modern Multidimensional Scaling: Theory and Applications. Springer, New York, 1997.

Table 1. Projection errors, obtained using Strategy II

p \ k	10	20	30	40	50
5	0.3039	0.2640	0.2532	0.2499	0.2468
10	0.3045	0.2572	0.2469	0.2482	0.2445
15	0.2991	0.2576	0.2459	0.2448	0.2430
20	0.3038	0.2555	0.2446	0.2434	0.2413
25	0.2982	0.2568	0.2436	0.2427	0.2408

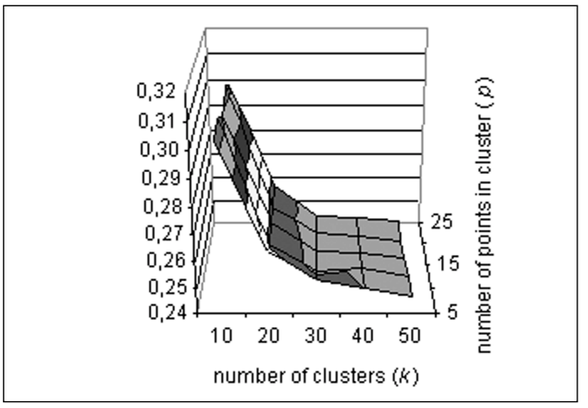


Fig 5. Surface of projection errors, obtained using Strategy II

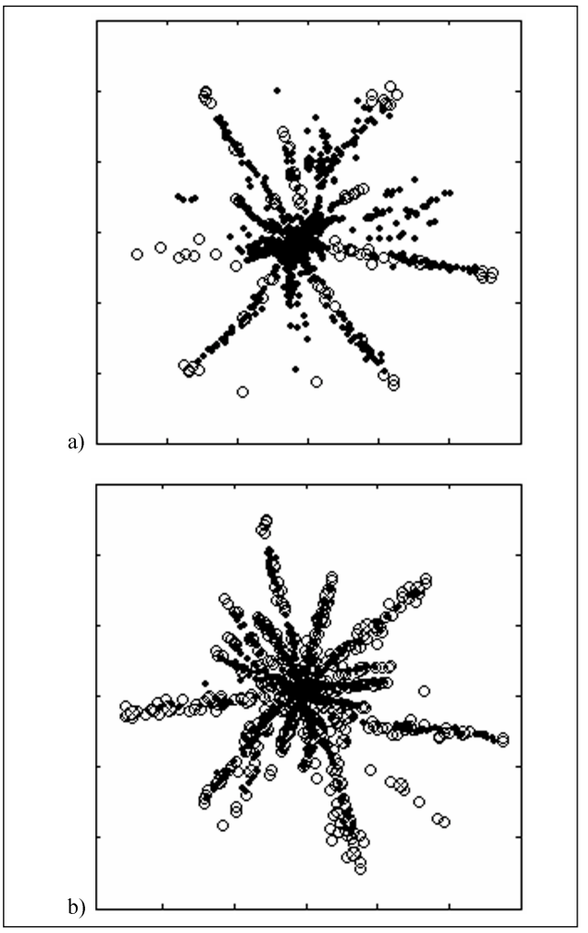


Fig 6. Projection of the ellipsoidal data: a) E =0.30360, b) E =0.24408

4. Kohonen, T. Self-Organizing Maps, third ed. *Springer Series in Information Sciences*, Vol 30, 2001.
5. Dzemyda, G.; Kurasova, O. Heuristic Approach for Minimizing the Projection Error in the Integrated Mapping. *European Journal of Operational Research*, Vol 171–3, 2006, p. 859–878.
6. Mao, J.; Jain, A. K. Artificial neural networks for feature extraction and Multivariate data projection. *IEEE Transactions on Neural Networks*, Vol 6, No 2, 1995, p. 296–317.
7. Podlipskyte, A. Visualization of multidimensional data and its application to biomedical data analysis. Summary of doctoral dissertation. Kaunas University of Technology. Kaunas, 2004. 34 p.
8. Pekalska, E.; de Ridder, D.; Duin, R. P.; Kraaijveld, M. A. A new method of generalizing Sammon mapping with application to algorithm speed-up. In: *Proceedings of 5th Annual Conference of the Advanced School for computing and image (ASCI'99), edited by Boasson M.; Karndorp J.; Torino J.; Vosselman M. 1999.*
9. Basalaj, W. Incremental multidimensional scaling method for database visualization. In: *Proceedings of Visual Data Exploration and Analysis VI*, SPIE, Vol 3647, 1999, p. 149–158.
10. Naud, A.; Duch, W. Interactive data exploration using MDS mapping. In: *Proceedings of the Fifth Conference: Neural Networks and Soft Computing*, 2000, p. 255–260.
11. Tiping, M. E. Topographic mappings and feed-forward neural networks. Ph. D. thesis. Aston University, Birmingham, UK, 1996.
12. Naud, A. Visualization of high-dimensional data using association of multidimensional scaling to clustering. In: *Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems*, Vol 1, 2004, p. 252–255.
13. Mathar, R.; Zilinskas, A. On Global Optimization in Two-Dimensional Scaling. *Acta Aplicandae Mathematicae,* Vol 33, 1993, p. 109–118.
14. Groenen, P. J. F.; van de Vaelden, M. Multidimensional Scaling. *Econometric Institute Report EI2004-15*, 2004, https://ep.eur.nl/handle/1765/1274/1/ei200415.pdf
15. Dunham, M. H. Data Mining Introductory and Advanced Topics, Pearson Education, Inc. (Prentice Hall), 2003.
16. Handl, J.; Knowles, J. Cluster generators for large high-dimensional data sets with large numbers of clusters. http://dbkgroup.org/handl/generators/

BAZINIŲ VEKTORIŲ PARINKIMO STRATEGIJŲ ANALIZĖ, TAIKANT SANTYKINĮ DAUGIAMAČIŲ SKALIŲ METODĄ

J. Bernatavičienė, G. Dzemyda, O. Kurasova, V. Marcinkevičius

Santrauka

Nagrinėjamas daugiamačių skalių metodas (MDS), pritaikytas didelių duomenų aibių analizei. Bendra algoritmo schema išskiriama į tris etapus: suformuojama bazinių vektorių aibė, paskui, naudojant klasikinį MDS algoritmą, baziniai vektoriai projektuojami į plokštumą, likusi duomenų aibė vizualizuojama, naudojant santykinį MDS algoritmą. Originaliame algoritme bazinių vektorių aibė formuojama, atsižvelgiant į k vidurkių klasterizavimo rezultatus. Šiame straipsnyje pasiūlytos dvi naujos bazinių vektorių parinkimo strategijos: vienoje taip pat atsižvelgiama į k vidurkių klasterizavimo rezultatus, tačiau kitu būdu, kitoje baziniais vektoriais parenkami duomenų aibės taškai. Eksperimentiniai tyrimai parodė, kad pasiūlytų strategijų naudojimas pagerina vizualizavimo kokybę, sutaupo skaičiavimo laiką.

Reikšminiai žodžiai: daugiamačių skalių metodas, vizualizavimas, klasterizavimas, bazinių vektorių aibė, naujų taškų vaizdavimas.

Jolita BERNATAVIČIENĖ. Doctoral student. Engineer programmer, Department of Systems Analysis, Institute of Informatics and Mathematics (IMI). First degree (higher education) in Mathematics and Informatics (1996), Master's degree in Informatics (2004) from Vilnius Pedagogical University. Research interests: visualization of multidimensional data, estimation of the visualization quality, estimation of data parameters.

Gintautas DZEMYDA. Professor, Doctor Habil, Department of Systems Analysis, Institute of Informatics and Mathematics (IMI). Doctor's degree in Technical Sciences in 1984 after post-graduate studies at IMI. Doctor Habil from Kaunas University of Technology in 1997. Research interests: interaction of optimisation and data analysis, optimisation theory and applications, multiple criteria decisions, neural networks, and data analysis.

Olga KURASOVA. Doctor of Sciences. Researcher, Department of Systems Analysis, Institute of Informatics and Mathematics (IMI). Lecturer, Department of Information Technology, Vilnius Pedagogical University. Doctor's degree in Informatics from IMI in 2005. Co-author of about 20 scientific articles. Research interests: visualization of multidimensional data, neural networks, self-organizing maps, clustering and classification.

Virginijus MARCINKEVIČIUS. Doctoral student. Engineer programmer, Department of Systems Analysis, Institute of Informatics and Mathematics (IMI). Bachelor's degree in Mathematics and Informatics (2001), Master's degree in Mathematics (2003) from Vilnius Pedagogical University. Research interests: neural networks, data mining, parallel computing.