ISSN 1392 - 1215   ELEKTRONIKA IR ELEKTROTECHNIKA. 2003. Nr. 6(48)

T 121 SIGNALŲ TECHNOLOGIJA

Investigation of Sequential Mapping of Multidimensional Data

A. M. Montvilas
Institute of Mathematics and Informatics, A. Goštauto 12, LT- 2600 Vilnius, Lithuania; e-mail: montvila@ktl.mii.lt
Vilnius Gediminas Technical University, Naugarduko 41, LT-2600 Vilnius, Lithuania;

1. Introduction

The *simultaneous* nonlinear mapping of data from *L*-dimensional space to a lower-dimensional space was created by Sammon [1]. The inherent structure of the data is approximately preserved under the mapping. It is achieved by minimizing the error function *E*, which reveals the largest product of the error and partial error [2].

$$E = \frac{1}{\sum\limits_{i<j}^{N} d_{ij}^{*}} \sum_{i<j}^{N} \frac{\left(d_{ij}^{*} - d_{ij}\right)^2}{d_{ij}^{*}}; \qquad (1)$$

where *N* is a number of *L*-dimensional vectors being mapped, d_{ij}^{*} - distance between *i* and *j* vectors in *L*-space, d_{ij} - distance in a lower-dimensional space (two-space).

The heuristic relaxation method [3] runs faster and requires a less amount of memory space. However, these two methods work only having all the data, already.

For the *sequential* nonlinear mapping the triangular method was presented by Lee [4]. It preserves only two distances to vectors previously mapped and, in addition, it uses the spanning tree, so it makes the mapping dependent on the history, hence it is usable only for very special tasks.

The *sequential* nonlinear mapping [5] occur to be very successful for sequential mapping of multidimensional data into a lower-dimensional space (frequently onto the plane). It can be applied either for the sequential clustering or for other sequential multidimensional data structure analysis or for supervising of dynamical systems, when each stable state of the system is described by parameters vector [6]. It allows us to watch the dynamical system states, their change and to indicate its damage [7].

In [8] this sequential method and Sammon's simultaneous one were compared according ability to map the data onto the plane, mapping accuracy and a mapping time. It was showed that sequential nonlinear mapping has slightly bigger total mapping error but needs incomparably less calculation time, and it was recommended to use the sequential nonlinear mapping for data structure analysis, even having all the data already, especially when there is a large amount of the data.

However, this sequential method needs some investigations.

In this paper the method of *sequential* nonlinear mapping has been investigated: a) according ability to differ the data groups when at the beginning the number of groups is taken to be less than really exists; b) according mapping errors dependence on a value of ***F*** ("magic factor") and, c) according mapping errors dependence on a sort of initial conditions.

2. Ability to Differ the Data Groups

The essence of any nonlinear mapping is to preserve the inherent structure of distances among the parameter's vectors being in *L*-dimensional space after mapping them into two-dimensional space. The *sequential* nonlinear mapping [5] requires at the very beginning to map *M* initial vectors *simultaneously*, using Sammon's algorithm. After that each sequentially receiving vector has to be mapped with respect to the first *M* vectors. Mapping error function E_j has to be minimized for each receiving vector *j*=*M*+1,…, *M*+*N*, using formula

$$E_j = \frac{1}{\sum\limits_{i=1}^{M} d_{ij}^{X}} \sum_{i=1}^{M} \frac{\left(d_{ij}^{X} - d_{ij}^{Y}\right)^2}{d_{ij}^{X}}, \quad j = M+1, \ldots, M+N; \quad (2)$$

where d_{ij}^{X} - distance between *i* and *j* vectors in the *L*-space, d_{ij}^{Y} - distance on the plane.

The set of the initials vectors *M* usually consists of the representatives of either stable state describing parameters vector of a dynamic system [6] or each cluster [8]. In other words the *M* initial vectors represent each of M^{*} data groups being mapped. Of course some times may be situations when *M* is not equal to M^{*}. It is not trouble, when $M>M^{*}$, but it is not clear how the mapping would behave if $M<M^{*}$.

To determine the mapping behavior, when $M<M^{*}$, a plenty of experiments has been executed. Two of most characteristic ones are presented. Let's have 25 vectors consisting of six parameters (Table 1). Let's call these data as "Data 1".

Table 1. 25 vectors of the "Data 1"

Ve ct. No	Parameters					
	1	2	3	4	5	6
1	50.0	220.0	307.0	508.0	604.0	206.0
2	51.0	222.0	302.0	501.0	603.0	205.0
3	53.0	228.0	304.0	502.0	607.0	208.0
4	55.0	226.0	305.0	505.0	609.0	204.0
5	56.0	224.0	308.0	507.0	608.0	202.0
6	50.3	220.0	307.0	507.7	604.0	206.0
7	50.1	220.2	307.2	508.0	604.0	206.3
8	50.4	220.7	307.0	508.2	604.0	206.1
9	50.2	220.5	307.1	507.9	603.9	205.8
10	51.6	222.2	302.0	501.0	603.0	205.1
11	50.9	221.8	302.1	500.9	603.1	204.9
12	50.7	222.4	302.2	501.2	603.2	205.0
13	50.8	222.6	302.2	501.5	603.0	205.5
14	52.5	228.4	304.2	502.0	607.0	208.0
15	52.9	227.8	304.2	502.3	607.3	208.1
16	53.2	228.2	304.0	502.0	607.0	208.2
17	53.1	227.9	303.8	501.8	606.4	207.9
18	55.0	226.5	304.5	504.7	608.7	203.9
19	55.1	225.9	305.5	504.9	609.1	204.1
20	54.9	226.2	304.9	505.0	609.2	204.0
21	55.2	225.8	304.8	505.2	609.0	204.0
22	56.2	224.2	307.9	507.0	609.2	202.0
23	55.9	224.4	308.4	507.5	608.0	202.5
24	56.3	223.8	309.0	507.5	607.2	202.2
25	55.8	224.1	308.6	507.0	608.0	201.9

According the experiment these vectors belong to five classes (Table 2)

Table 2. 25 vectors of "Data 1"distributed to five classes

Class	Vectors
1	1,6,7,8,9,
2	2,10,11,12,13,
3	3,14,15,16,17,
4	4,18,19,20,21
5	5,22,23,24,25

For this investigation it was used 100 iterations R. In Fig. 1 the result of sequential mapping vectors into two-dimensional space at M=4 is presented. The first M=4 vectors mapped simultaneously are denoted by mark x with an index that means the vector's number, and the remainder N=20 vectors mapped sequentially are denoted by mark + with the respective index. The representative of fifth class was not involved into M vectors, however, all vectors of fifth class were mapped into separate group with some distances to other groups.

Now let's execute the sequential mapping using the same "Data 1" at M=3.
In the Fig. 2 the result of mapping the these data at M=3 is presented.

Neither fifth nor fourth classes representatives were not involved into M vectors. This case the classes were separated as well, nevertheless, the distance between fourth and fifth classes is smaller than other distances. In Table 3 the mapping total errors are presented at M=5,4 and 3. The mapping total errors were calculated including distances among *all* vectors like in the case of simultaneous mapping (formula (1)).

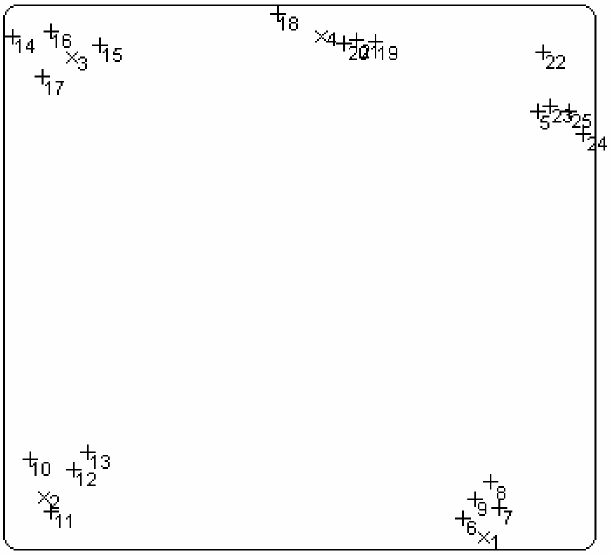

Fig. 1. Mapping of the "Data 1" at M=4

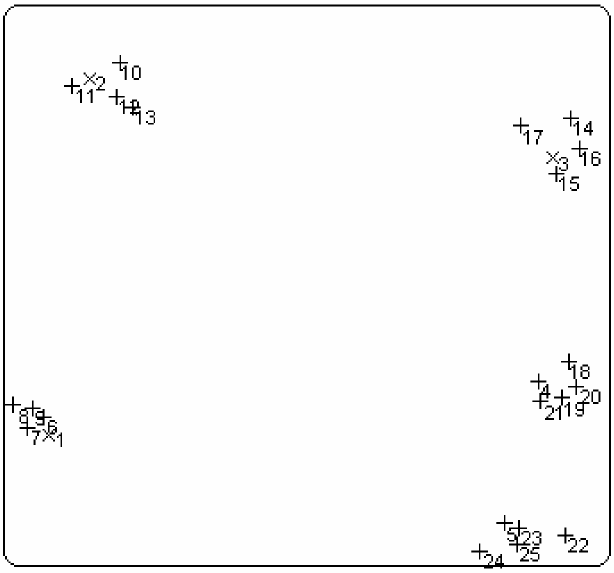

Fig. 2. Mapping of the "Data 1" at M=3

Table 3. Mapping total errors of "Data 1"

M	Total errors
5	0.005383999
4	0.007153901
3	0.012378900

Let's repeat the similar mapping using another kind of data ("Data 2"). In the Table 4 the "Data 2" consisting of 30 vectors is presented. Each vector consists of six parameters, as well.

The result of sequential mapping of the "Data 2" at M=4 is presented in Fig. 3 and at M=3 in Fig.4 respectively.

The mapping total errors at M=5,4 and 3 are presented in the Table 6.

They are distributed to five classes as well (see Table 5):

Table 4. 30 vectors of the "Data 2"

Vec.	Parameters					
No.	1	2	3	4	5	6
1	1.0	5.0	7.0	8.0	4.0	6.0
2	6.0	3.0	2.0	1.0	3.0	5.0
3	8.0	2.0	4.0	2.0	7.0	8.0
4	3.0	8.0	5.0	5.0	9.0	4.0
5	5.0	3.0	4.0	7.0	8.0	2.0
6	2.5	7.5	5.1	4.8	9.0	3.0
7	3.0	7.8	5.5	4.9	9.1	4.1
8	2.9	7.9	4.9	5.0	9.2	4.0
9	3.1	8.1	4.8	5.2	9.0	4.0
10	5.2	3.3	5.0	7.2	7.5	2.0
11	5.1	3.2	4.8	7.1	7.8	2.1
12	4.8	3.1	4.2	6.8	7.7	2.2
13	3.1	8.7	5.3	5.2	9.3	4.1
14	3.3	8.3	5.5	5.5	9.8	4.3
15	3.2	8.4	5.4	5.1	8.2	4.4
16	8.1	2.1	4.0	2.1	7.1	8.3
17	9.0	2.0	4.2	2.0	7.0	8.0
18	8.0	3.0	4.8	2.3	7.0	8.0
19	5.4	2.1	4.0	7.0	7.0	2.0
20	7.9	1.8	3.8	1.0	6.4	7.9
21	7.0	1.5	3.9	1.8	6.8	8.0
22	6.0	4.5	2.1	0.9	3.1	6.0
23	6.0	3.0	2.9	1.2	3.2	5.0
24	6.1	3.2	2.3	1.2	3.2	4.1
25	6.2	3.1	1.9	1.2	3.2	4.1
26	6.1	2.9	1.8	1.0	3.0	4.0
27	1.0	5.1	7.0	7.6	4.0	6.0
28	1.0	5.2	8.0	8.0	4.0	6.3
29	1.0	5.1	7.0	7.8	4.0	6.0
30	1.1	5.5	6.0	9.0	4.0	6.0

Table 5. 30 vectors of "Data 2" distributed to five classes

Class	Vectors
1	1,27,28,29,30
2	2,22,23,24,25,26
3	3,16,17,18,20,21
4	4,6,7,8,9,13,14,15
5	5,10,11,12,19

Table 6. Mapping total errors of "Data 2"

M	Total errors
5	0.02750241
4	0.03240145
3	0.07661533

Analysing the mapping results we see that at M=4 the mapping procedure differs the data into the groups well. However, at M=3 the first three data groups are separated well meantime the reminder data (of fourth and fifth groups) are mapped into more or less one place. The mapping total error increasing by decreasing M (Table 3 and Table 4) shows the deteriorate of mapping quality. A great deal of experiments have been executed with various sort of data.

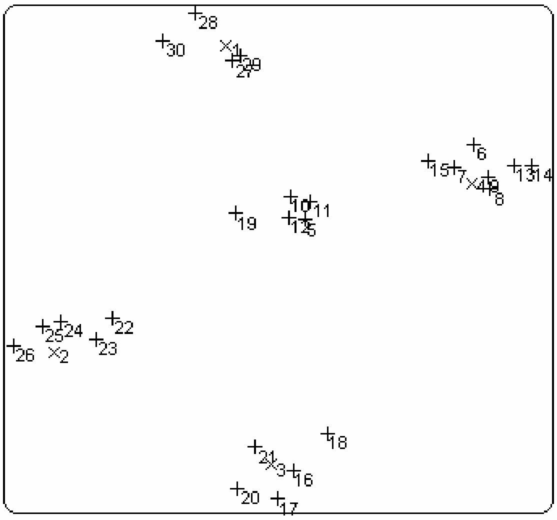

Fig. 3. Mapping of the "Data 2" at M=4

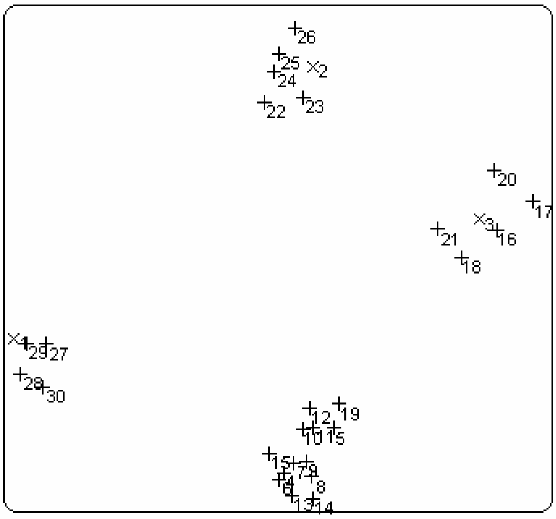

Fig. 4. Mapping of the "Data 2" at M=3

Always groups which representatives were involved into M vectors were mapped correctly, and the reminder data were separated into give or take close place on the screen. So, if the number of groups M^* is not known, the sequential nonlinear mapping is more or less correct when the number of the initial vectors M is taken to be at least $M=M^*$-1.

3. Mapping Errors Dependence on the Value of *F*

F-factor in [5], or *MF*-"magic factor" in [1] is using for correction of the vector's co-ordinates on the plane during each iteration procedure. In [9] it was proposed to take the ***F*** value between 0.3 and 0.4 and in [10] between 0.25 and 0.45. However, investigations carried out show that for the sequential nonlinear mapping the range of ***F*** could be taken more wide. A lot of experiments show that mapping error's curve by changing ***F*** in the range from 0.05 to 1.0 has rather wide part with minimum mapping error, and only at the ends of the range the mapping errors grow up. The range of ***F*** with minimum mapping error and minimum error value depend on the nature of data and the number of iterations. For illustration that, the mapping error both the first 5 vectors and sequential average and total sequential for "Data 1" and "Data 2" at 50 and 200 iterations are presented in Fig.5 – Fig.8, respectively. The errors were calculated at several sort of initial conditions and then averaged.

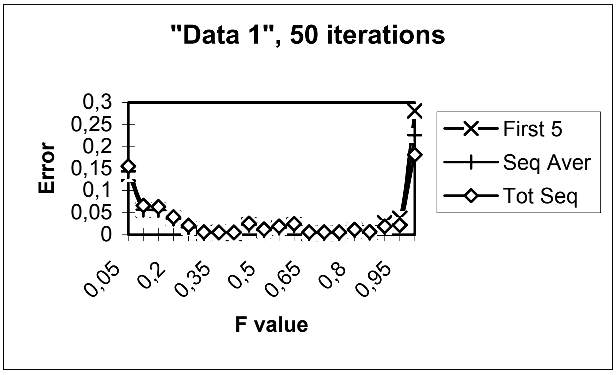

Fig. 5. Dependence the mapping errors on ***F*** for "Data 1" at 50 iterations

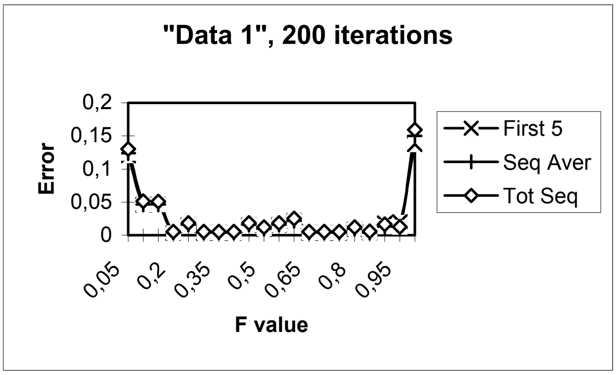

Fig. 6. Dependence the mapping errors on ***F*** for "Data 1" at 200 iterations

A great deal of experiments have been executed using various sort of data and experiment's conditions. They showed that the factor ***F*** for correction of the co-ordinates on the plane for the sequential nonlinear mapping can be taken in the range from 0.25 to 0.75.

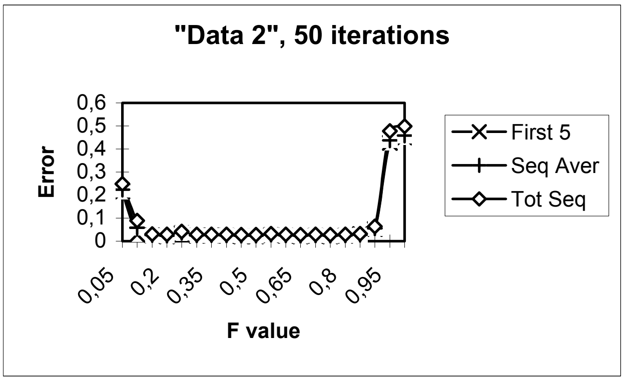

Fig. 7. Dependence the mapping errors on ***F*** for "Data 2" at 50 iterations

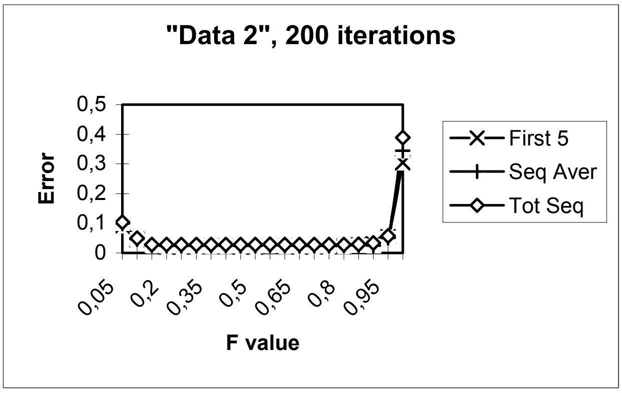

Fig. 8. Dependence the mapping errors on ***F*** for "Data 2" at 200 iterations

4. Mapping Error Dependence on Initial Conditions

Mapping error depends on the nature of data, but it especially depends on initial conditions because any nonlinear mapping algorithm often finds the local maximum of a functional that characterises the mapping quality which is not global [11] because of bad initial conditions. There are many ways to choose initial conditions using certain knowledge of the data. The matter is that most often we have only the data without any knowledge about it. This general case initial points on the plane are distributed along more or less shifted diagonal.

The experiments have been carried out using several kinds of slightly shifted descending or ascending diagonals:

D - descending diagonal,
DS - descending shifted diagonal,
DMS - descending more shifted diagonal,
DOMS - descending one more shifted diagonal,
A - ascending diagonal,
AS - ascending shifted diagonal,
AMS - ascending more shifted diagonal,
AOMS - ascending one more shifted diagonal.

The mapping error of the first *M* vectors, the average of sequential mapping errors, the sequential mapping total error along with Sammon's mapping error were calculated. In Fig. 9 the results at 500 iterations are presented for "Data 1" and in Fig. 10 for "Data 2", respectively.

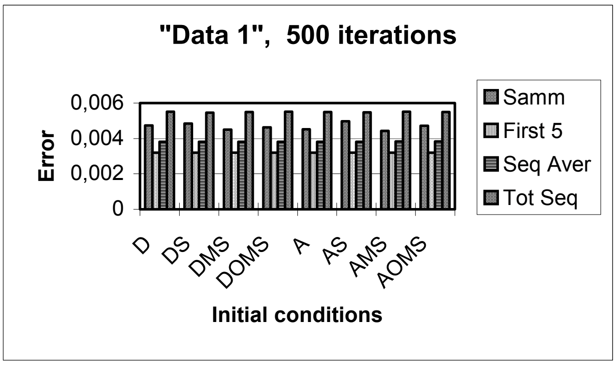

Fig. 9. Magnitudes of the errors at various initial conditions for "Data 1"

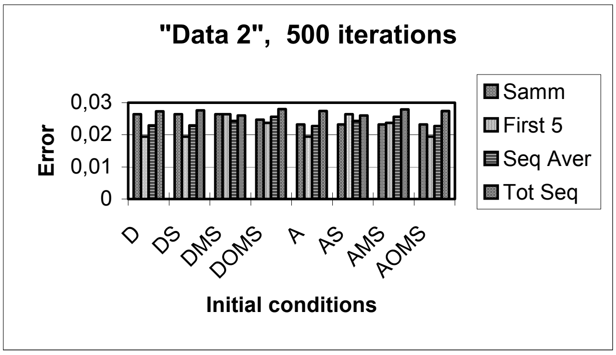

Fig. 10. Magnitudes of the errors at various initial conditions for "Data 2"

The experiments show that for the "Data 1" every mapping error is of a little value and they change only a little by changing the initial conditions. For the "Data 2", on the contrary, the errors are rather bigger than that of the "Data 1", and their values change itself by changing the initial conditions in a more extent value. It means that for the "Data 1" this sort of initial conditions is almost optimal and the mapping quality functional is near to the global maximum.

Numerous experiments, have been executed using various sort of data, show that mapping errors depend on both the sort of initial conditions and the nature of data. Notice that comparative values of four kinds of errors mentioned above remain give or take the same.

5. Conclusions

The method of the sequential nonlinear mapping has been investigated according ability to differ the data groups when the number of groups M is taken less than really exists M^*, according mapping errors dependence on a value of the factor F for correction of co-ordinates on the plane and according mapping errors dependence on a sort of initial conditions.

It was showed that the sequential nonlinear mapping differs the groups of data when the number of initial vectors M, mapped simultaneously, is taken to be at least $M=M^*-1$. The factor F for correction co-ordinates on the plane for the sequential nonlinear mapping can be taken in the range from 0.25 to 0.75. Mapping errors depend on both the sort of initial conditions and the nature of data.

References

1. **Sammon J.W.** A nonlinear mapping for data structure analysis // *IEEE Trans. on Computers.*- 1969.- Vol. c-18(5).- P. 401-409.
2. **Duda R.O., P.E. Hart.** *Pattern Classification and Scene Analysis.*-New York, London, Sydney, Toronto: John Wiley & Sons, 1973.
3. **Chang C. L., R. C. T. Lee**. A heuristic relaxation method for nonlinear mapping in cluster analysis // *IEEE Transactions on Systems, Man and Cybernetics*.- 1973.- **3.-** P. 197-200.
4. **Lee R.C.T., J.R. Slangle, H. Blum.** A triangulation method for the sequential mapping of points from *N*-space to *two*-space // *IEEE Trans. on Computers*.- 1977.-Vol. c-26(3).-P. 288-292.
5. **Montvilas A.M.** On sequential nonlinear mapping for data structure analysis // *Informatica.*- 1995.-6(2).-P. 225-232.
6. **Montvilas A.M.** Issues for design of information system for supervision and control of dynamic systems // *Informatica.*- 1999.-**10**(3).- P. 289-296.
7. **Montvilas A. M.** Processing of information for supervision and control of technological processes // *Proceedings of the IFAC Workshop*.- 2000.-Vienna: Pergamon press.-P. 39-43.
8. **Montvilas A. M.** Sequential nonlinear mapping versus simultaneous one // *Informatica*.-2002.- 13(3).-P. 333-343.
9. **Kohonen T.** *Self-Organizing Maps*// 3nd ed. *Springer Series in Information Sciences.*-Springer-Verlag.- 2001.-Vol. 30.
10. **Groenen P. J. F., W. J. Heiser.** Tunnelling method for global optimisation in multidimensional scaling // *Psychometrica*.-1996.- 61.- P. 529-550.
11. **Dzemyda G.** Clustering of parameters on the basis of correlations: a comparative review of deterministic approaches // *Informatica*.-1997.- 8(1).-P. 83-118.

Pateikta spaudai 2003 05 30

A. M. Montvilas. Nuoseklaus daugiamačių duomenų atvaizdavimo tyrimas // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2003. – Nr. 6(48) – P. 7-12.

Pateikti nuoseklaus netiesinio daugiamačių duomenų atvaizdavimo metodo [5], taikomo duomenų struktūrai analizuoti bei vizualizuoti, tyrimų rezultatai. Šis metodas skiriasi nuo Sammono vienalaikio daugiamačių duomenų atvaizdavimo į plokštumą metodo [1] tuo, kad po pradiniame etape vienu metu atvaizduotų keleto duomenų vektorių vėliau galima dirbti nuosekliai realiu laiku. Šiuo būdu galima stebėti technologinių procesų arba dinaminių sistemų būsenas, aprašomas daugelio parametrų vektoriumi, stebėti būsenų pasikeitimus bei matyti momentinius sistemų gedimus [7]. Metodas buvo tiriamas pagal gebėjimą skirstyti duomenis į grupes (klasterizuoti), kai duomenų grupių skaičius pradžioje imamas mažesnis, negu yra iš tikrųjų, pagal atvaizdavimo klaidos priklausomybę nuo koordinačių plokštumoje koregavimo koeficiento dydžio bei pagal atvaizdavimo klaidos priklausomybę nuo pradinių sąlygų ir duomenų tipo. Parodyta, kad nuoseklaus netiesinio atvaizdavimo metodu duomenys gerai skirstomi į grupes net kai pradinis duomenų grupių skaičius imamas vienetu mažesnis už tikrą; koordinačių plokštumoje koregavimo koeficientą nuosekliam atvaizdavimui galima imti nuo 0,25 iki 0,75; atvaizdavimo klaida priklauso nuo nedidelio pradinių sąlygų pakeitimo bei nuo duomenų tipo. Il.10, bibl.11 (anglų kalba; santraukos lietuvių, anglų ir rusų k.).

A. M. Montvilas. Investigation of Sequential Mapping of Multidimensional Data // Electronics end Electrical Engineering. – Kaunas: Technologija, 2003. – No. 6(48) – P. 7-12.

In the paper the results of investigations of the sequential nonlinear mapping [5] of multidimensional data for data structure analysis and visualization are presented. This method differs from Sammon's method of simultaneous nonlinear mapping of multidimensional data onto the plane [1] that after simultaneous mapping of several data vectors at the very beginning, later one can work sequentially in a real time. This way one can watch the states of technological processes or dynamic systems, which are described by vectors consisting of many parameters, watch state's changes and see instant system's failures [7]. The method has been investigated according ability to differ the data groups (clustering) when at the beginning the number of data groups is taken to be less than really exists, according mapping errors dependence on a value of factor for correction co-ordinates on the plane and according mapping errors dependence on initial conditions. It was showed that the sequential nonlinear mapping differs the groups of data when the number of initial vectors, mapped simultaneously, is taken to be less by one than really exists. The factor for correction co-ordinates on the plane for the sequential nonlinear mapping can be taken in the range from 0.25 to 0.75. Mapping errors depend on both the sort of initial conditions and the nature of data. Ill. 10, bibl. 11 (in Lithuanian; summary in Lithuanian, English and Russian).

А. М. Монтвилас. Исследование последовательного отображения многомерных данных // Электроника и электротехника. – Каунас: Технология, 2003. – № 6(48) – С. 7-12.

В работе представлены результаты исследования метода последовательного отображения многомерных данных [5] для анализа структуры данных и визуализации. Этот метод отличается от метода Саммона одновременного отображения многомерных данных на плоскости [1] тем, что в начале, после одновременного отображения всего нескольких векторов данных, далее можно работать последовательно, в реальном времени. При этом можно следить за состояниями технологических процессов или динамических систем, описываемыми векторами многих параметров, следить за изменениями состояний и мгновенно фиксировать выход из строя систем [7]. Метод был исследован по способности разделения данных по группам (кластеризации), когда в начальном отображении количество групп принимается меньшим, чем есть на самом деле, по зависимости ошибки отображения от величины коэффициента коррекции координат на плоскости и по зависимости ошибки отображения от начальных условий и типа данных. Показано, что метод последовательного нелинейного отображения хорошо разделяет данные по группам даже когда в начале количество групп принимается на единицу меньше, чем в действительности; для последовательного отображения величина коэффициента коррекции координат может быть от 0,25 до 0,75; ошибка отображения зависит от незначительного изменения начальных условий и от типа данных. Ил.10, библ. 11 (на английском языке; резюме на литовском, английском и русском яз.).