ELECTRONICS AND ELECTRICAL ENGINEERING

ISSN 1392 – 1215 *2009. No. 1(89)*

ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY

T 120

SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

Dependence of Sammon and Sequential Mapping Errors on Data Configuration

A. Montvilas

Electronic Systems Department, Faculty of Electronics, Vilnius Gediminas Technical University, Naugarduko str. 41, LT-03227 Vilnius, Lithuania, e-mail: algirdas.montvilas@gmail.com

Introduction

There are many methods for visualization of multidimensional data by mapping them into three or two-dimensional space. The purpose of mapping is to preserve the inherent structure of distances among the vectors in *L*-dimensional space after mapping them onto the plane. A more precise structure of distances is preserved using nonlinear mapping methods. The classical Sammon [1] mapping is used for map the whole data simultaneously. In order to supervising the states and their changes of complicated dynamic systems the sequential nonlinear mapping has been created [2]. This sequential mapping requires at the very beginning to map only several (*M*) initial vectors using Sammon mapping. Thereafter each sequentially receiving vector of parameters has to be mapped in respect to the first *M* initial vectors. By the way, the *M* vectors should represent all the possible system states or clusters. In any case the mapping errors are inevitable. Consequently it is necessary to minimize them in any way. Frequently it is necessary to map the data as exactly as possible without respect to computing expenses.

Both Sammon method and sequential one uses the iteration procedure for calculating the vector's co-ordinates on the plane by minimizing mapping error. The steepest-descent method is used for that. Overall, mapping error depends on many conditions because the steepest-descent algorithm often finds the local maximum of a functional that characterizes the mapping quality which is not global [3]. Dependence of mapping error on initial conditions has been investigated in [4]. In [5] sequential mapping results was proposed to take as initial conditions for Sammon mapping. In most cases the Sammon mapping was improved (mapping error become less). Usually data array has any unpredictable constitution. So it is very important to know, how mapping errors depend on parameters scattering and data array configuration. Such investigations were not found in any references.

In the paper both Sammon and sequential nonlinear mapping has been investigated according their mapping errors dependence on scatter of parameters and data array configuration.

Peculiarity of Sammon and sequential mapping algorithms

Before mapping, two-dimensional vectors on the plane are distributed in a various way (randomly, diagonal, ring, et cetera). Mapping error E_S for Sammon method is calculated by formula which reveals the largest product of error and partial error [6]:

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

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

The sequential mapping begins after mapping the M vectors X_i, i=1,…, M, simultaneously using Sammon algorithm. The last procedure is only a preparing for sequential mapping (the first stage). Mapping error for sequential method E_j, $j=M+1,\ldots, M+N$, is calculated for each sequentially receiving vector X_j, $j=M+1,\ldots, M+N$:

$$\begin{cases} E_j = \dfrac{1}{\sum\limits_{i=1}^{M} d_{ij}^X} \sum\limits_{i=1}^{M} \dfrac{\left(d_{ij}^X - d_{ij}^Y\right)^2}{d_{ij}^X}, \\ j = M+1,\ldots, M+N, \end{cases} \qquad (2)$$

where N – a number of sequentially receiving vectors, d^X_{ij} – distance between i and j vectors in the L – space, d^Y_{ij} – corresponding distance on the plane.

After mapping whole data array sequentially the total mapping error by formula (1) can be calculated for comparison. The sequential mapping often gives the total mapping error a bit bigger than that of Sammon one [7], but it enables us to map *on-line* and thereby to watch dynamic system's states and even suddenly indicate systems false or damage.

During each iteration r both Sammon algorithm and sequential one corrects co-ordinates of the vectors on the plane by:

$$y_{pq}(r+1) = y_{pq}(r) - F * \Delta_{pq}(r), \qquad (3)$$

where p=1,…,N; q=1,2; coefficient F is called "magic factor" (0.3-0.4 in [8] or 0.25-0.45 in [9]);

$$\Delta_{pq}(r) = \frac{\partial E(r)}{\partial y_{pq}(r)} \Bigg/ \left| \frac{\partial^2 E(r)}{\partial y^2_{pq}} \right|. \qquad (4)$$

The magnitude of second derivation of $E(r)$ in (4) makes impossible to research mapping methods analytically. Therefore there is the only way to investigate the methods using simulation.

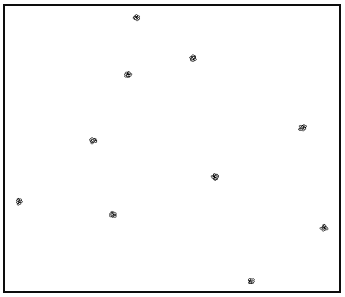
In the following experiments mapping errors have been calculated by formula (1) in three cases:

- for Sammon mapping (S1);
- for sequential mapping (SQ);
- for Sammon mapping using sequential mapping results as initial conditions (S2).

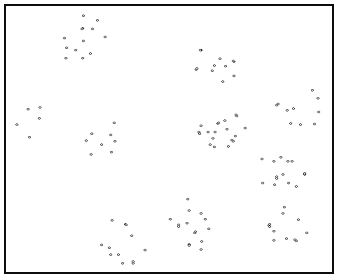
Influence of parameter's scatter on mapping error

In reality parameters of vectors may be of various scattering. Some data gives definite distribution of clusters (Fig. 1a). Another data have a scattering of parameters, and there is no evidence about clusters or system states (Fig. 1c). The mapped data (number of vectors N=120, dimension L=10, number of clusters M=10) in three cases of parameters scattering level are shown in Fig. 1: a) scatter level 0; b) scatter level 5; c) scatter level 10.

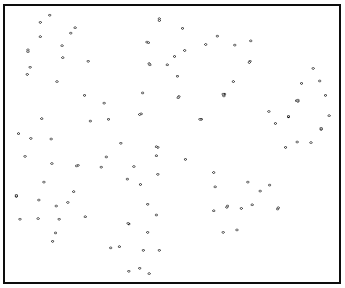
A lot of experiments have been executed to evaluate the influence of parameter's scattering over mapping errors. A typical experiment is given below with the data: number of vectors N=120, dimensionality L=10, number of clusters M=10, scattering of parameters: S=0,1,…,10. Dependence of mapping errors of three cases: S1, SQ, S2, on parameters scattering level, from 0 to10, are presented in the Table 1, and corresponding diagram is shown in Fig. 2.



(a)



(b)



(c)

Fig. 1. View of mapped data: N=120, L=10, M=10; (a)-scatter label 0, (b)-scatter label 5, (c)-scatter label 10

Table 1. Dependence of mapping errors on scatter of parameters

Scatter level	S1	SQ	S2
0	0.066901	0.056338	0.051335
1	0.067205	0.065641	0.048711
2	0.065856	0.065112	0.060145
3	0.061561	0.066067	0.061559
4	0.067129	0.074330	0.064887
5	0.050614	0.057181	0.050005
6	0.082963	0.093422	0.082842
7	0.065665	0.070414	0.065612
8	0.068901	0.089369	0.068765
9	0.072316	0.082330	0.072326
10	0.081538	0.089277	0.081538

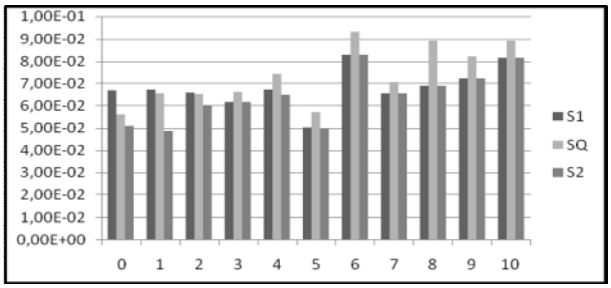


Fig. 2. Dependence of three cases of mapping errors on parameters scatter

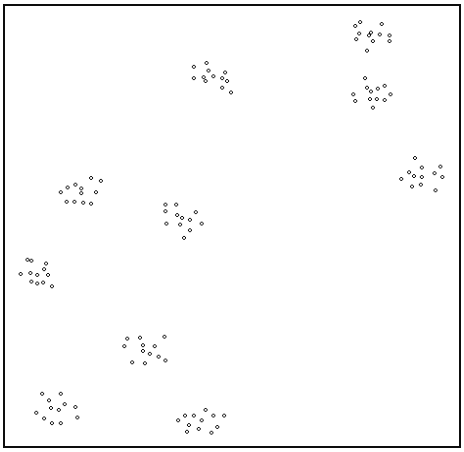
Many executed experiments give similar results. One can draw inference from plenty experiments that parameters scatter weekly influence on mapping error, unless the sequential mapping error is less than that of Sammon one at a small scatter of parameters. Other experiments give more interesting result.

Data array configuration influence on mapping error

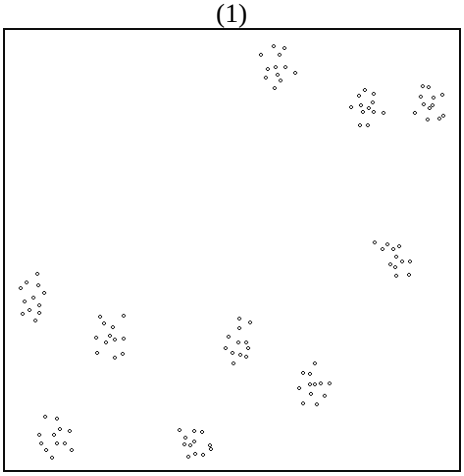
Usually we have data array for mapping as matrix of various numbers, whichever represents parameters of vectors being mapped. If we map data array simultaneously (Sammon method) it seems that any array configuration fit equally, and every other configuration will give the same mapping results (the same pattern and mapping error). But actually next experiments show that this assumption is wrong.

On the other hand, sequential method required at the first stage to map *M* vectors simultaneously. Usually these vectors are the first *M* vectors of data array being mapped. Of course a pattern and total mapping error (formula (1)) depend on which *M* vectors were used for the first stage.

Data array has been recomposed by changing their matrix rows groups amongst and building in such a way several new configurations of data array.



(1)



(2)

Fig. 3. View of mapping result of two configurations of the same data

The mapping patterns of two different configurations of the same data are shown in Fig. 3. Data array consisted of *N*=120 vectors of only 6 parameters and had 10 clusters. Sammon method was used.

A big number of experiments had been executed for investigation the influence of various configuration of data array on mapping error. As an example we take data array matrix consisting of 120 rows (vectors). Each vector is described by 6 parameters, and data consists of 10 clusters. Mapping errors (S1, SQ, S2) depending on data array configurations (12) are presented in the Table 2, and corresponding diagram in Fig. 4.

Table 2. Mapping errors of various mapping and data configurations

Configu-ration	S1	SQ	S2
1	0.014491	0.015401	0.014473
2	0.021866	0.023596	0.021790
3	0.027518	0.031014	0.027456
4	0.014482	0.016271	0.014491
5	0.018915	0.021879	0.018892
6	0.027468	0.033848	0.027462
7	0.014496	0.017329	0.014475
8	0.014493	0.016698	0.014482
9	0.014473	0.016571	0.014475
10	0.014468	0.016689	0.014491
11	0.019591	0.022595	0.019597
12	0.014487	0.017737	0.014491

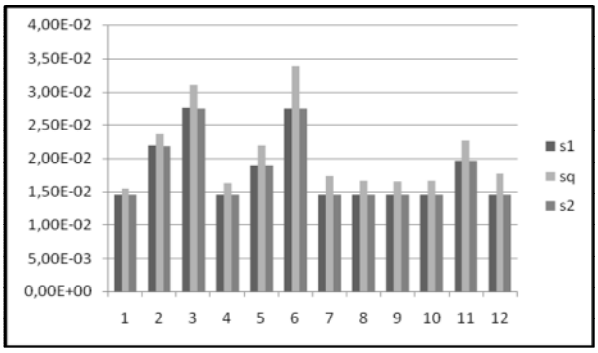


Fig. 4. Dependence of mapping errors on data array configuration

Experiments show that while mapping the same data but of various configuration the mapping errors change noticeably (almost twice).

Conclusions

1. Both Sammon mapping error end sequential one weakly depend on scattering of parameters of vectors being mapped, but mapping errors noticeably change at various configurations of data array.

2. In order to get the best mapping (the smallest mapping error) it is recommended to map the data at

several various configuration of the data array and choose the best one. It is an alternative way instead of changing the initial conditions.

3. Sequential mapping often gives smaller mapping errors at some data array configuration than that of Sammon one at another data array configuration.

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. **Montvilas A. M.** On sequential nonlinear mapping for data structure analysis // Informatica. – 1995. – No. 6(2). – P. 225–232.
3. **Dzemyda G.** Clustering of parameters on the basis of correlations: a comparative review of deterministic approaches // Informatica. – 1997. – No. 8(1). – P. 83–118.
4. **Montvilas A. M.** Features of sequential nonlinear mapping // Informatica. – 2003. – No. 14(3). – P. 337–348.
5. **Montvilas A. M.** Optimal initial conditions for nonlinear mapping of multidimensional signals // Electronics and Electrical Engineering. – Kaunas: Technologija, 2005. – No. 1(57). – P. 24–27.
6. **Duda R. O., Hart P. E., and Stork D. G.** Pattern Classification, 2nd Edition. – John Wiley & Sons. – 2000.
7. **Montvilas A. M.** Sequential nonlinear mapping versus simultaneous one // Informatica. – 2000. – No. 13(3). – P. 333–343.
8. **Kohonen T.** Self-Organizing Maps, 3rd ed. – Springer Series in Information Sciences. – Springer-Verlag, 2001. – Vol. 30.
9. **Groenen P. J. F., and Heiser W. J.** Tunnelling method for global optimisation in multidimensional scaling // Psychometrica. – 1996. – No. 61. – P. 529–550.

Received 2008 06 05

A. Montvilas. Dependence of Sammon and Sequential Mapping Errors on Data Configuration // Electronics and Electrical Engineering. – Kaunas: Technologija, 2009. – No. 1(89) – P. 25–28.

The Sammon nonlinear mapping along with the sequential nonlinear mapping has been investigated according their mapping errors dependence on data structure and on scatter of parameters of vectors in multidimensional space. A lot of experiments show that mapping error weakly depends on scattering of parameters but noticeably depends on configuration of data array. Examples are given. Ill 4, bibl. 9 (in English; summaries in English, Russian and Lithuanian).

А. Монтвилас. Зависимость ошибки Саммона и последовательного отображения от конфигурации данных // Электроника и электротехника. – Каунас: Технология, 2009. – № 1(89). – С. 25–28.

Исследуется зависимость ошибок отображения Саммона и последовательного отображения от разброса параметров в многомерном пространстве и от конфигурации массива данных. Многочисленные эксперименты показывают, что ошибка отображения мало зависит от разброса параметров, но в значительной степени зависит от конфигурации массива данных. Приведены примеры. Ил. 4, библ. 9 (на английском языке; рефераты на английском, русском и литовском яз.).

A. Montvilas. Sammono ir nuoseklaus atvaizdavimo klaidos priklausomybė nuo duomenų konfigūracijos // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2009. – Nr. 1(89). – P. 25–28.

Nagrinėjama daugiamačių vektorių parametrų sklaidos ir duomenų masyvo struktūros įtaka Sammono vienalaikio netiesinio atvaizdavimo bei nuoseklaus netiesinio atvaizdavimo klaidoms. Parodyta, kad netiesinio atvaizdavimo klaidos mažai priklauso nuo parametrų sklaidos laipsnio daugiamatėje erdvėje, bet labai priklauso nuo duomenų masyvo konfigūracijos. Pateikta pavyzdžių. Il. 4, bibl. 9 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).