ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 *2006. Nr. 1(65)*
ELEKTRONIKA IR ELEKTROTECHNIKA

T 125 *AUTOMATION, ROBOTICS*
AUTOMATIZAVIMAS, ROBOTOTECHNIKA

Data Structure Influence on Mapping Error

A. M. Montvilas
Institute of Mathematics and Informatics, A. Goštauto str. 12, LT- 01108 Vilnius, Lithuania; e-mail:montvila@ktl.mii.lt
Vilnius Gediminas Technical University, Naugarduko str. 41, LT-03227 Vilnius, Lithuania

Introduction

Multidimensional data are mapped onto the plane for their visualization. The Sammon's method of nonlinear mapping often is used for this purpose [1]. Multidimensional data are described by many parameters, so they are presented in L-dimensional space, where L-number of parameters. The essence of the method is to preserve the inner structure of distances among the vectors in multidimensional space after mapping them into two-dimensional space. Mapping error is calculated by:

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

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).

Formula (1) reveals the largest product of error and partial error [2], and works effectively at both big distances and small ones, therefore it is used as criteria for mapping quality.

For the first, vectors on the plane, as initial conditions, are distributed randomly or along diagonal. Formula (2) is used for correction of co-ordinates of the vectors on the plane during each iteration r.

$$y_{pq}(r+1) = y_{pq}(r) - F * \Delta_{pq}(r), \qquad (2)$$

where p=1,…,N; q=1,2; F is "*magic factor"* (0.3-0.4);

$$\Delta_{pq}(r) = \frac{\partial E(r)}{\partial y_{pq}(r)} \Bigg/ \left| \frac{\partial^2 E(r)}{\partial y_{pq}^2(r)} \right|. \qquad (3)$$

The Sammon's method maps multidimensional data onto the plane simultaneously. In order to watch mapped data *on-line* a method of sequential nonlinear mapping has been created [3]. It allowed us to watch dynamic systems states and their changes [4]. The essence of sequential mapping is that at the first stage several data M are mapped onto the plane using Sammon's method, and at the second stage sequentially receiving data are mapped with respect to simultaneously mapped the initial M vectors. Mapping error function E_j for each receiving vector X_j, l=M+1,…,M+N is calculated using formula:

$$E_j = \frac{1}{\sum_{i=1}^{M} d_{ij}^{X}} \sum_{i=1}^{M} \frac{\left(d_{ij}^{X} - d_{ij}^{Y}\right)^2}{d_{ij}^{X}}, \quad j = M+1,..., M+N; \quad (4)$$

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 [3] or each cluster.

The mapping error (1) for both Sammon's and the first stage of sequential mapping depends on initial conditions, on what manner points on the plane are distributed initially (before iteration procedure), because any nonlinear mapping algorithm often finds the local maximum of a functional that characterizes the mapping quality which is not global [5].

Influence of initial conditions and other factors on mapping error has been investigated in [6,7]. There is no possibility to avoid the mapping error. In the paper data structure influence on mapping error is investigated.

Influence of data structure on mapping quality

Multidimensional vectors being mapped could be distributed to some clusters. Let such data be denominated as regular data. On the other hand, vectors parameters could be random and submitted to e.g. uniform or normal law. Let these data be named as random data. One can have several or many vectors, which could consist of few or many parameters. Investigations have been executed considering all factors mentioned above. A great deal of experiments has been executed. Several typical experiments are presented in the paper.

For the first the parameters of vectors that belong to M=10 clusters were generated: N=120, L=12. Unfortunately it is no possibility to present here this big matrix of the data: 12x120. Notice that parameters were from the values of big three-figure number to small two-figure number. Mapping result of these data using R=500 iterations is presented in Fig. 1. The mapping error was E=0.02444897.

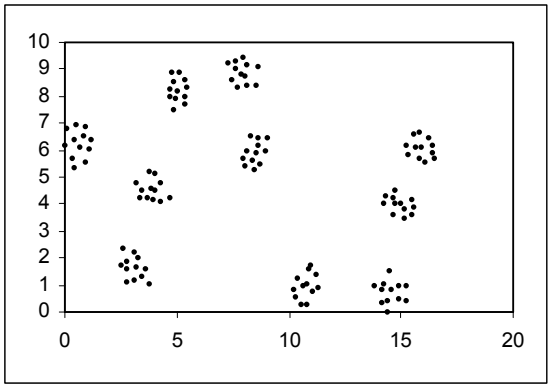

Fig. 1. Mapping result of 120 vectors consisting of 12 parameters and belonging to 10 clusters

Mapping errors have been calculated at number of parameters L=3,4,…,12 and at number of vectors N=20,40,…,120; in all 60 times. Dependence of error E on number of parameters L is presented in Fig. 2. Mapping errors were averaged by all calculated numbers of vectors N.

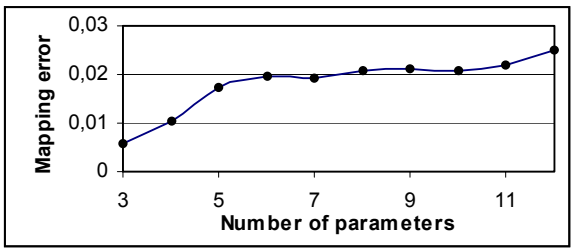

Fig. 2. Dependence E on L averaged by N for regular signal

Dependence of mapping error E on number of vectors N averaged by all numbers of parameters L is presented in Fig. 3.

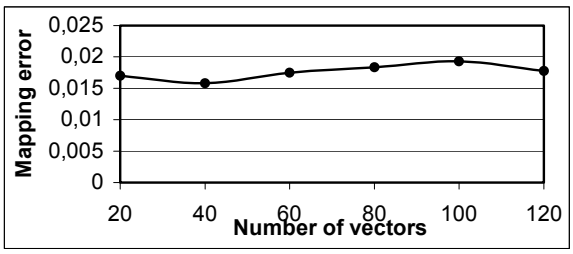

Fig. 3. Dependence E on N averaged by L for regular signal

Another experiments have been executed using random numbers as parameters of vectors. 1440 random numbers have been generated from the values of 0 to 1 *uniformly* distributed. Distribution is showed in Fig. 4.

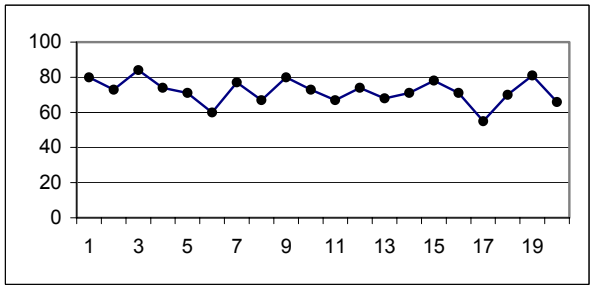

Fig. 4. Uniform distribution of random parameters

They composed the data matrix 12x120 as well. Mapping result of N=120 vectors is presented in Fig. 5. The mapping error was E=0.1267093.

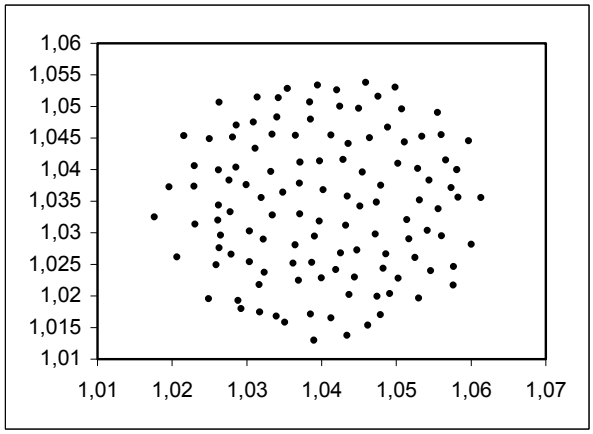

Fig. 5. Mapping result of 120 vectors whose parameters are uniformly distributed random numbers

Mapping errors have been calculated in the same way like of previous experiment with 10 clusters (60 times). Dependence of mapping error E on number of parameters L averaged by all numbers of vectors N is presented in Fig. 6.

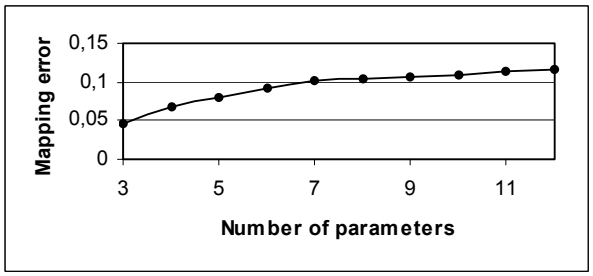

Fig. 6. Dependence E on L averaged by N for uniformly distributed random parameters

Dependence of mapping error E on number of vectors N averaged by L is presented in Fig. 7.

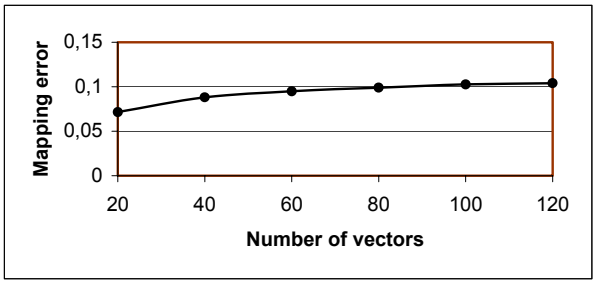

Fig. 7. Dependence E on N averaged by L for uniformly distributed random parameters

Finally, experiment similar to previous one has been executed using *normal* distributed random numbers from the values of 0 to 1 having average 0.5 and standard deviation 0.15. Distribution is shown in Fig. 8.

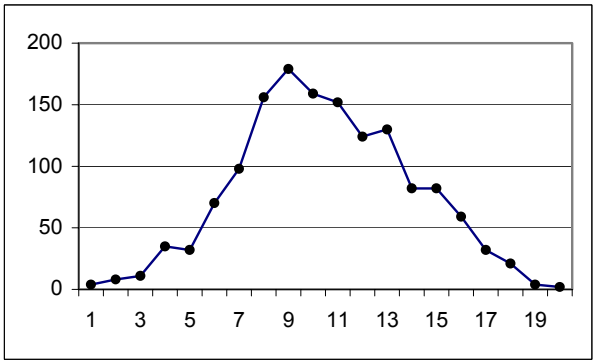

Fig. 8. Normal distribution of random parameters

Mapping result of N=120 vectors with normal distributed parameters is presented in Fig. 9. The mapping error was E=0.1887024.

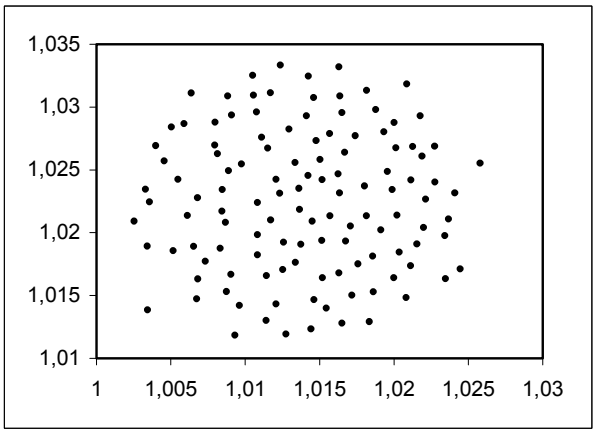

Fig. 9. Mapping result of 120 vectors with normal distributed parameters

Dependence of mapping error E on number of parameters L averaged by all numbers of vectors N for normal distributed random parameters is shown in Fig. 10.

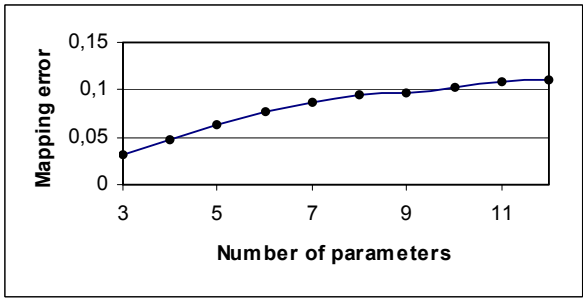

Fig. 10. Dependence E on L averaged by N for normally distributed random parameters

Dependence of mapping error E on number of vectors N averaged by all numbers of parameters L for normal distributed random parameters is presented in Fig.11.

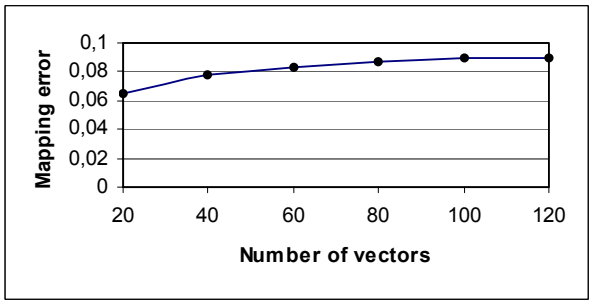

Fig. 11. Dependence E on N averaged by L for normally distributed random parameters

Experiments with various another data having regular or random parameters give results similar to them presented above. Every time mapping error E changes several times by changing number of parameters L and changes a little by changing number of vectors N many times.

Conclusions

1. Independent on regular or random data:
   - mapping error E increases, increasing number of vector's parameters L,
   - mapping error E remains almost constant, increasing number of vectors N.
2. Vectors with random parameters are distributed after mapping them onto the plane in similar way, independent on what order parameters are distributed: *uniform* or *normal*.
3. Mapping errors of regular data are several times less than that of random 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., Hart P.E.** Pattern Classification and Scene Analysis. – New York, London, Sydney, Toronto: John Wiley & Sons, 1973.
3. **Montvilas A.M.** On Sequential Nonlinear Mapping for Data Structure Analysis // Informatica.–1995.–**6**(2).–P. 225-232.
4. **Montvilas A.M.** Issues for design of information system for supervision and control of dynamic systems // Informatica.–1999.–**10**(3). – P. 289–296.
5. **Dzemyda G.** Clustering of Parameters on the Basis of Correlations: a Comparative Review of Deterministic Approaches // Informatica.–1997.– **8**(1).–P. 83–118.
6. **Montvilas A. M.** Investigation of Sequential Mapping of Multidimensional Data // Electronics and Electrical Engineering. – Kaunas: Technologija, 2003.–No.6(48).–P.7–12.
7. **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.

Pateikta spaudai 2005 05 30

A. M. Montvilas. Duomenų struktūros įtaka atvaizdavimo klaidai // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2006. – Nr. 1(65) – P. 34–37.

Daugiamačiams duomenims atvaizduoti plokštumoje naudojamos įvairios procedūros. Labai paplitęs yra Sammono vienalaikio netiesinio daugiamačių duomenų atvaizdavimo plokštumoje metodas. Daugiamačiams duomenims stebėti realiu laiku sukurtas nuoseklaus atvaizdavimo metodas. Atvaizdavimo esmė – išlaikyti vidinę atstumų tarp duomenų vektorių struktūrą po jų atvaizdavimo. Atvaizdavimo kokybę nusako atvaizdavimo klaida, kuri priklauso nuo pradinių sąlygų ir nuo duomenų struktūros. Darbe ištirta atvaizdavimo klaidos priklausomybė nuo duomenų struktūros. Gausybė eksperimentų, keičiant vektorių parametrų kiekį bei duomenų vektorių kiekį, atvaizduojant reguliarius (turinčius informaciją apie klasterius) bei atsitiktinius duomenis, leidžia daryti išvadas, kad, didėjant parametrų kiekiui, atvaizdavimo klaida didėja, didėjant vektorių kiekiui, atvaizdavimo klaida keičiasi nedaug. Be to, reguliarių duomenų atvaizdavimo klaida yra kelis kartus mažesnė negu atsitiktinių duomenų atvaizdavimo klaida. Il.11, bibl.7 (anglų kalba; santraukos lietuvių, anglų ir rusų k.).

A. M. Montvilas. Data Structure Influence on Mapping Error // Electronics end Electrical Engineering. – Kaunas: Technologija, 2006. – No. 1(65) – P. 34–37.

Various procedures are used for mapping of multidimensional data onto the plane. Sammon's method of simultaneous nonlinear mapping onto the plane is very popular. In addition, sequential nonlinear mapping has been created for watching the data in real time. The essence of mapping is to preserve the inner structure of distances among the vectors in multidimensional space after mapping them onto the plane. The mapping error characterizes the mapping quality, and it depends on initial conditions and data structure. The paper deals with investigations of how data structure influences on mapping quality. Plenty experiments have been executed at various number of parameters and various number of vectors using regular data (having information about clusters) and random data. The experiments allowed to draw conclusions that mapping error increases increasing the number of parameters, and it remains almost constant increasing the number of vectors. Besides, mapping error of regular data is several times less than that of random data. Ill. 11, bibl. 7 (in Lithuanian; summaries in Lithuanian, English and Russian).

А. М. Монтвилас. Влияние структуры данных на ошибку отображения // Электроника и электротехника. – Каунас: Технология, 2006. – № 1(65) – С. 34–37.

Для отображения многомерных данных на плоскости используются различные процедуры. При этом широко используется метод Саммона одновременного нелинейного отображения векторов данных на плоскости. Для наблюдения многомерных данных в реальном времени был разработан метод последовательного отображения. Суть отображения заключается в том, чтобы сохранить внутреннюю структуру расстояний между многомерными векторами после их отображения на плоскости. Качество отображения характеризует ошибка отображения, которая зависит от начальных условий и от структуры данных. В работе исследуется влияние структуры данных на качество отображения. Многочисленные эксперименты при различном количестве параметров и самих векторов, как регулярных данных (несущих информацию о кластерах), так и случайных данных, позволяет делать выводы, что ошибка отображения увеличивается при увеличении количества параметров и меняется незначительно при увеличении количества векторов. Кроме того, ошибка отображения регулярных данных несколько раз меньше, чем ошибка отображения случайных данных. Ил.11, библ. 7 (на английском языке; резюме на литовском, английском и русском яз.).