INFORMATICA, 1995, Vol. 6, No. 2, 225–232

# ON SEQUENTIAL NONLINEAR MAPPING FOR DATA STRUCTURE ANALYSIS

Algirdas Mykolas MONTVILAS

Institute of Mathematics and Informatics
2600 Vilnius, Akademijos St. 4, Lithuania

**Abstract.** An algorithm for the sequential analysis of multivariate data structure is presented. The algorithm is based on the sequential nonlinear mapping of $L$-dimensional vectors from the $L$-hyperspace into a lower-dimensional (two-dimensional) vectors such that the inner structure of distances among the vectors is preserved. Expressions for the sequential nonlinear mapping are obtained. The mapping error function is chosen. Theoretical minimum amount of the very beginning simultaneously mapped vectors is obtained.

**Key words:** dimensionality reduction, sequential nonlinear mapping, sequential detection of changes.

**1. Introduction.** The purpose of this paper is to present essential investigations of the sequential nonlinear mapping algorithm which has been found to be highly effective in the sequential analysis of multivariate data structure. Technological process or dynamic system (further – DS) can be described either by various parametric models, consisting of $L$ parameters: $a_1, a_2, \ldots, a_L$, or by random process generated by this DS. Then the DS state is represented by $L$ data characterizing the random process too. When the DS state changes the $L$ parameters describing the DS change as well. The DS can have several unknown states and we need to identify the states and to detect their changes sequentially and independently of the history. It is convenient to watch the DS states and their changes marking them by some mark on PC screen and, having in mind the existence of particular states, to identify the current state, a deviation from it or a transition to other state when the mark changes its position.

For solution of this problem it is necessary to have a method of sequential detection of many changes in several unknown properties of random processes.

There are many methods of detection of changes in the properties of random processes in the scientific publications (Kligienė and Telksnys, 1984; Basseville and Benveniste, 1986; Nikiforov, 1983), but there are no methods to solve the above mentioned problems, because DS can have several unknown states, the DS states can change themselves abruptly or slowly and we need to watch many changes of DS states sequentially and in a real time. In Montvilas (1993) the sequential nonlinear mapping for data analysis is presented and it has been found to be effective in sequential analysis of DS. This method is based on the sequential nonlinear mapping onto the plane of vectors of the $L$ parameters representing the DS states. In this paper the sequential nonlinear mapping is considered, the mapping error function is chosen and theoretical minimum amount of at the very beginning simultaneously mapped vectors is obtained. The last is of great theoretical and practical importance, because it let us avoid theoretically possible mistakes of sequential mapping when the amount of the first $M$ simultaneously mapped vectors of parameters is taken to be the smallest ($M = 2$).

**2. Sequential nonlinear mapping.** Let a DS be in any state $s_i$ of the set of possible states: $s_i \in S$. We can watch vector of $L$ parameters at the output of the DS. If these parameters are of different physical nature, we must introduce the scale coefficients for each parameter. For identification of states of the DS or detection their changes it is necessary at discrete time moments to map the $L$-dimensional vectors sequentially and nonlinearly into two-dimensional vectors in order to reflect the present state by some mark on the PC screen. According to the mark position on the screen we can make a decision of DS state and its abrupt or slow change if the mark position changes. The main requirement of mapping the $L$-dimensional vectors into two-dimensional vectors is to preserve the inner structure of distances among the vectors. This is achieved using a nonlinear mapping procedure.

A sequential mapping requires existence of earlier on mapped vectors, so at the very beginning we have to carry out the nonlinear mapping of the $M$ vectors ($M \geqslant 2$) simultaneously. The expressions in Sammon (1969) were used for that. Afterwards we map sequentially and nonlinearly the receiving parameter vectors, and, in such a way, we can identify the states and detect their various changes for a practically unlimited time. In order to formalize the method we denote by $N$ this practically unlimited number of the arriving vectors.

Now we present the sequential nonlinear mapping algorithm and choose the mapping error function. Let us have $M + N$ vectors in the $L$-hyperspace. We denote them $X_i$, $i = 1, \ldots, M$; $X_j$, $j = M + 1, \ldots, M + N$. The $M$ vectors are already simultaneously mapped into two-dimensional vectors $Y_i$, $i = 1, \ldots, M$, using expressions in Sammon (1969). Now we need to map sequentially the $L$-dimensional vectors $X_j$ into two-dimensional vectors $Y_j$, $j = M+1, \ldots, M+N$. Here the simultaneous nonlinear mapping expressions will change into sequential nonlinear mapping expressions, respectively. First, before performing iterations it is expedient to put the two-dimensional vectors being mapped in the same initial conditions, i.e. $y_{jk} = C_k$, $j = M + 1, \ldots, M + N$; $k = 1, 2$. Note, that in the case of simultaneous mapping of the first $M$ vectors the initial conditions are chosen in a random way (Sammon, 1969). Let the distance between the vectors $X_i$ and $X_j$ in the $L$-hyperspace be defined by $d^X_{ij}$ and on the plane – by $d^Y_{ij}$, respectively. This algorithm uses the Euclidean distance measure, because, if we have no a priori knowledge concerning the data, we would have no reason to prefer any metric over the Euclidean metric (Sammon, 1969).

For computing the mapping error of distances $E$ we can find at least three expressions.

$$E_1 = \frac{1}{\sum_{i=1}^{M} (d^X_{ij})^2} \sum_{i=1}^{M} (d^X_{ij} - d^Y_{ij})^2, \quad j = M + 1, \ldots, M + N; \qquad (1)$$

function $E_1$ reveals the largest errors independently of magnitudes of $d^X_{ij}$; but if $d^X_{ij}$ is small then the mapping error can be comparable with the same distance.

$$E_2 = \sum_{i=1}^{M} \left( \frac{d^X_{ij} - d^Y_{ij}}{d^X_{ij}} \right)^2, \quad j = M + 1, \ldots, M + N; \qquad (2)$$

function $E_2$ reveals the largest partial errors independently of magnitudes of $|d^X_{ij} - d^Y_{ij}|$; but in this case big distance will have rather big mapping error.

$$E_3 = \frac{1}{\sum_{i=1}^{M} d^x_{ij}} \sum_{i=1}^{M} \frac{(d^X_{ij} - d^Y_{ij})^2}{d^X_{ij}}, \quad j = M + 1, \ldots, M + N; \qquad (3)$$

function $E_3$ is the useful compromise and reveals the largest product of error and partial error. So we choose the third expression for computing the mapping error of distances $E$.

For correct mapping we have to change the positions of vectors $Y_j$, $j = M+1, \dots, M+N$ on the plane in such a way that the error $E$ would be minimal. This is achieved by using the steepest descent procedure. After the $r$-th iteration the error of distances will be

$$E_j(r) = \frac{1}{\sum_{i=1}^{M} d_{ij}^X} \sum_{i=1}^{M} \frac{\left[d_{ij}^X - d_{ij}^Y(r)\right]^2}{d_{ij}^X}, \qquad (4)$$
$$j = M+1, \dots, M+N,$$

where

$$d_{ij}^Y(r) = \sqrt{\sum_{k=1}^{2} \left[y_{ik} - y_{jk}(r)\right]^2}, \qquad (5)$$
$$i = 1, \dots, M, \; j = M+1, \dots, M+N.$$

During the $r+1$ iteration the coordinates of the mapped vectors $Y_j$ will be

$$y_{jk}(r+1) = y_{jk}(r) - \boldsymbol{F}\Delta_{jk}(r), \qquad (6)$$
$$j = M+1, \dots, M+N; \; k = 1, 2,$$

where

$$\Delta_{jk}(r) = \frac{\partial E_j(r)}{\partial y_{jk}(r)} \Bigg/ \left| \frac{\partial^2 E_j(r)}{\partial y_{jk}^2(r)} \right|. \qquad (7)$$

$\boldsymbol{F}$ is the coefficient for correction of the coordinates and it is defined empirically to be $\boldsymbol{F} = 0.35$;

$$\frac{\partial E}{\partial y_{jk}} = H \sum_{i=1}^{M} \frac{D \cdot C}{d_{ij}^X d_{ij}^Y}, \qquad (8)$$

$$\frac{\partial^2 E}{\partial y_{jk}^2} = H \sum_{i=1}^{M} \frac{1}{d_{ij}^X d_{ij}^Y} \left[ D - \frac{C^2}{d_{ij}^Y} \left( 1 + \frac{D}{d_{ij}^Y} \right) \right], \qquad (9)$$

where

$$H = -\frac{2}{\sum_{i=1}^{M} d_{ij}^x}, \quad D = d_{ij}^X - d_{ij}^Y, \quad C = y_{jk} - y_{ik}.$$

After some iterations the error of distances will be $E_j < \varepsilon$, where $\varepsilon$ can be taken arbitrary small, the iteration process is over and result is shown on the PC screen. In fact it is enough $\varepsilon = 0.01$. In order to have equal computing time for each mapping we can execute constant number of iterations $R$. In practice it is enough $R = 30$.

**3. Minimum necessary value of the $M$.** While executing experiments, in all generated situations, when amount of at the very beginning initial simultaneously mapped vectors of parameters of DS states was taken to be $M = 2$, at every time moments the marks of DS states got into their right places on the PC screen, and the states were identified correctly (Montvilas, 1992; 1993; 1994). Even marks of those states which were not involved into $M$ initial vectors of parameters had got their own places on PC screen and at every time moment the places of marks on the screen corresponded to right DS states entirely.

However, theoretically there are possible such cases, when points (ends of parameter vectors) being in different places in the $L$-hyperspace can be mapped into one point on the plane, because these points have the same distances with $M = 2$ simultaneously mapped points. By way of illustration let us map points from three-dimensional space ($L = 3$) onto the plane. Let the initial simultaneously mapped points are $A$ and $B$ (see Fig. 1). Let they are being on the axis of cylinder. Then points $C$, $D$ and $E$ being on circle, which is on surface of cylinder, have the equal distances with point $A$ and point $B$: $d^X_{AC} = d^X_{AD} = d^X_{AE}$ and $d^X_{BC} = d^X_{BD} = d^X_{BE}$.

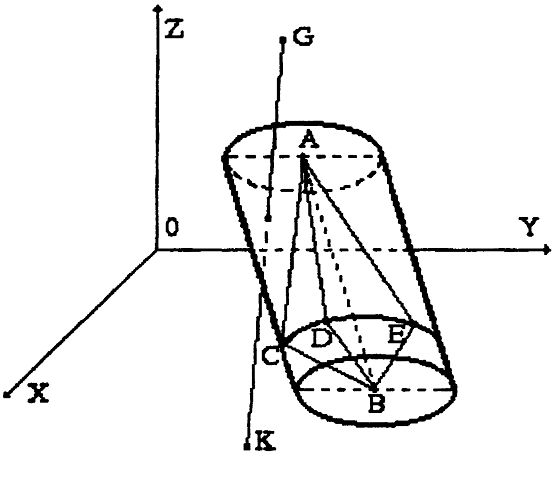

Fig. 1. The case of three-dimensional space ($L = 3$).

So points $C$, $D$ and $E$ having the same initial conditions will be mapped onto the plane into the same point. If points $C$, $D$ and $E$ reflect the different states of DS then we shall have a mistake.

No let us have three points $M = 3$ as the initial points for simultaneous mapping: $A$, $B$ and $C$. Then one can draw any straight line orthogonal to the plane $ABC$. Any two points of the line $G$ and $K$ being on different sides of the plane $ABC$ have equal distances with points $A$, $B$ and $C$: $d^X_{AG} = d^X_{AK}$, $d^X_{BG} = d^X_{BK}$ and $d^X_{CG} = d^X_{CK}$. So in this case points $G$ and $K$ can be mapped into one point, too.

When taking $M = 4$ points ($A$, $B$, $C$ and $D$) so that the fourth $D$ point would not be on the plane $ABC$ and all $M = 4$ points would form the three-dimensional space we have a situation when even theoretically one can not find any two points, which have equal distances with all simultaneously mapped $M = 4$ points.

Thus, having analysis of various possible situations at diverse $L$ and $S$ values, we can draw a conclusion that for initial simultaneous mapping one need to take $M = \min(L + 1, S)$ when $S$ is known or $M = L + 1$ when $S$ is unknown or dynamic system can have indeterminate or spoiled states, besides, these $M = L + 1$ points have to form the $L$-dimensional space.

In practice, how it was mentioned above, it is enough to take $M = 2$ vectors of DS states for the initial simultaneous mapping, because cases considered here can take place only under coincidence of unexpectedness. However, in order to avoid only theoretically possible complications, we need to do the following: after having simultaneous mapping of $M = 2$ and sequential mapping of $L - 1$ vectors we have got $L + 1$ vectors, already. Then we have to map simultaneously the available $L + 1$ vectors again and after that to map sequentially the receiving later vectors with respect to the initial $L + 1$ vectors.

**4. Conclusions.** The considered method of sequential nonlinear mapping of vectors of parameters from the $L$-hyperspace onto the plane enables us either to sequentially detect many abrupt or slow changes in several unknown properties of random processes or to sequentially identify the dynamic systems states, their jumpwise or slow changes and to watch the situation on the PC screen.

Before sequential identification of the states or detection of changes in properties of random processes, it suffices to map simultaneously only $M = 2$

state vectors. However, in order to avoid probably only theoretically possible complications one need to take $M = L + 1$, where $L$ is the dimensionality of vectors of parameters which describe the dynamic system states or random process properties. Besides, these $M = L + 1$ points have to form the $L$-dimensional space.

## REFERENCES

Basseville, M., and A.Benveniste (1986). *Detection of Abrupt Changes in Signals and Dynamical Systems*. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo.

Kligienė, N., and L.Telksnys (1984). Methods of detecting instants of change of random process properties. – A survey. *Automation and Remote Control,* **44**(10), 1241–1283.

Montvilas, A.-M. (1992). Sequential detection of many changes in several unknown states of dynamic systems. *Informatica,* **3**(1), 72–79.

Montvilas, A.-M. (1993). A sequential nonlinear mapping for data analysis. *Informatica,* **4**(1–2), 81–93.

Montvilas, A.-M. (1994). Discrete sequential detection of abrupt or slow multiple changes in several unknown properties of random processes. *Informatica,* **5**(1–2), 175–188.

Nikiforov, I. (1983). *Sequential Detection of Change in the Properties of Time Series.* Nauka, Moscow (in Russian).

Sammon, J.W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. on Computers,* **18**(5), 401–409.

Received October 1994

**A.-M. Montvilas** received the degree of Candidate of Technical Sciences from the Kaunas Technological University, Kaunas, Lithuania, in 1974. He is a senior researcher at the Process Recognition Department of the Institute of Mathematics and Informatics. Scientific interests include: processing of random processes, classification, detection of a change in the properties of random processes, simulation.

## NUOSEKLAUS NETIESINIO ATVAIZDAVIMO DUOMENŲ STRUKTŪROS ANALIZEI KLAUSIMU

Algirdas Mykolas MONTVILAS

Pateikiamas daugiamačių duomenų struktūros analizės algoritmas, kuris remiasi $L$-mačių vektorių, aprašančių duomenis, nuosekliu netiesiniu atvaizdavimu į dvimačius vektorius, išsaugant vidinę atstumų tarp jų struktūrą. Gautos nuoseklaus netiesinio atvaizdavimo išraiškos. Parinkta atvaizdavimo paklaidos funkcija. Gautas teorinis minimalus pačioje pradžioje vienalaikiai atvaizduojamų vektorių kiekis.