Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 12, Number 5 (2016), pp. 3817-3922
© Research India Publications
http://www.ripublication.com/gjpam.htm

# Neighborhood Factor Structures

**Anatoliy Shmyrin, Semyon Blyumin and Anastasia Kosareva**

*Department of Mathematics, Lipetsk State Technical University,
Lipetsk, 398600, Russia.*

## Abstract

We consider the clustering problem for graphs of a special type (neighborhood structures) provided with a set of experimental data and some natural metric.

**Keywords:** neighborhood structure; metric; clustering; factor structure.

## INTRODUCTION

The neighborhood structure (see [1]) is a convenient means of formalizing the inter-element connections of the modeled system. This structure may be considered as a "discrete frame" on the basis of which it is possible to build mathematical models of the higher level, e.g. linear or bilinear. In [1] mathematical models built above a neighborhood structure are called neighborhood systems. In modeling systems with a large number of elements it is necessary to build a factor structure containing a lesser number of aggregated elements (clusters). For this purpose it is necessary to define a metric on a set of system elements and a rule for transforming the connections between them into the connections between clusters.

## NEIGHBORHOOD STRUCTURES

We define a *neighborhood structure* as a weighed oriented graph with $n$ vertices and edges of two types which may be called $s$-edges and $u$-edges. These edges symbolize two types of connections between vertices: state connections and control connections. Two vertices $a_i$ and $a_j$ can be connected by not more than two $s$-edges (from $a_i$ to $a_j$ and from $a_j$ to $a_i$) and, similarly, by not more than two $u$-edges. It may be considered that two such $s$-edges (or $u$-edges) are one edge with two orientations: from $a_i$ to $a_j$ and from $a_j$ to $a_i$. Graph vertices are called *nodes* of a neighborhood structure. The weight of each edge is a pair of numbers $(\tilde{c}, \bar{c})$ from the interval $[0,1]$, so that $\tilde{c} + \bar{c} >$

$0$. We suppose that the edge with the weight $(\tilde{c},\bar{c})$, leading from the $a_i$ node to the $a_j$ node, describes two connections of $a_i$ with $a_j$: *a soft* connection with the weight $\tilde{c}$ and *a hard* connection with the weight $\bar{c}$. If $\bar{c}=0$, there is only a soft connection of nodes, if $\tilde{c}=0$, there is only a hard connection of nodes. There will be a correspondence between soft connections and *calculated* coefficients of a mathematical model, and between rigid connections and *given* coefficients.

A state (control) *neighborhood* of node $a_i$ is a set of all end nodes of all $s$-edges ($u$-edges) proceeding from node $a_i$. We will describe the neighborhood structure by the *adjacency matrix* $[\tilde{S},\bar{S},\widetilde{U},\overline{U}]$ comprising four $n\times n$ order matrixes. Row $\tilde{S}_i=[\tilde{s}_{i1},\dots,\tilde{s}_{in}]$ of matrix $\tilde{S}$ sets the weights of state soft (calculated) connections of the $a_i$ node: $\tilde{s}_{ij}\in(0{,}1]$ (connection weight) if the $a_j$ node enters the neighborhood of the $a_i$ node by state and $\tilde{s}_{ij}$=0 if the $a_j$ node does not enter this neighborhood. The rows of the remaining $\bar{S},\widetilde{U},\overline{U}$ matrixes are defined similarly. Usually $\tilde{s}+\tilde{s}_{ij}\neq 0$ and $\tilde{u}+\tilde{u}_{ij}\neq 0$, i.e. each node enters its neighborhoods by state and by control (this corresponds to the loops of the graph), but, generally speaking, we do not exclude the case when $\tilde{s}+\tilde{s}_{ij}=0$ or $\tilde{u}+\tilde{u}_{ij}=0$.

*Note*. All the edges of the graph in [1] are single in our terminology, which means that they describe either soft or hard connections. In our designations it corresponds to one or the other number $\tilde{c}$ or $\bar{c}$ being zero. The existence of weight multipliers for the connections of the $i$-node with the other nodes is treated in [1] as the *fuzziness* of the corresponding neighborhoods.

## DATA TUPLES

The statistics of observing a real system described by a neighborhood structure will be written in the form of a $K\times 2n$ data matrix $[X_1,\dots,X_n,V_1,\dots,V_n]$, where $K$ is the number of observations. Rows $[X^k,V^k]$ of the data matrix are the results of the $k$ observation (experiment): $X^k=[x_1^k,\dots,x_n^k]$ and $V^k=[v_1^k,\dots,v_n^k]$ - are states and controls in the nodes of the structure which were observed in the $k$ experiment. These rows will be called *data tuples*. Columns $X_i$ or $V_i$ of the data matrix are data vectors of all observations of the corresponding node. Generally, a state or control in the nodes of the structure can be of a multiparameter character and therefore the elements of the data matrix can be vectors (generally, of different dimensions). In the elementary case which we will be considering further, these elements are *numbers*. Thus, we will consider states and controls in the structure nodes as *scalars*. This restriction is not vital, and the further considerations can be rewritten for the case of vector states and controls.

*Note*. It will be considered that the units of measurement for states and controls in each node are selected in such a way that the samples $\{x_i^k\}$, $k=1,\dots,K$ and $\{v_i^k\}$, $k=1,\dots,K$ (here $i$ – the node number) are centered and normed. The transition to such units of measurement is always possible, but, as a rule, requires the recalculation of the given («rigid») coefficients of the mathematical model.

## NEIGHBORHOOD SYSTEMS

A most simple mathematical model associated with a neighborhood structure (and describing the functioning of the corresponding real system) is a *linear symmetric neighborhood system (model)*. Within our approach to defining the metric and to further clustering a neighborhood structure, it would be possible to consider more sophisticated models, e.g. bilinear ones (see [1]), but for the sake of simplicity only a linear case will be considered. Further, we will use the designation $A \circ B$ for the element-wise product (Hadamard product) of matrixes $A$ and $B$. The linear symmetric neighborhood system takes the form

$$\left(\widetilde{\Omega} \circ \tilde{S}\right)X + \left(\tilde{T} \circ \widetilde{U}\right)V = (\overline{\Omega} \circ \bar{S})X + (\bar{T} \circ \overline{U})V \qquad (1)$$

where $[\tilde{S}, \bar{S}, \widetilde{U}, \overline{U}]$ is the earlier defined $n \times 4n$ matrix of the neighborhood structure, $\widetilde{\Omega} = \{\widetilde{\omega}_{iq}\}$ and $\tilde{T} = \{\tilde{\tau}_{iq}\}$ is $n \times n$ - matrixes of the calculated coefficients of the system, $\overline{\Omega} = \{\overline{\omega}_{iq}\}$ and $\bar{T} = \{\bar{\tau}_{iq}\}$ is $n \times n$ - matrixes of the given coefficients of the system, $X$ and $V$ is $n$ -tuples of states and controls which are unknowns of the neighborhood system. The calculated coefficients of the system correspond to soft connections, the given coefficients correspond to hard connections. It should be reminded that we consider the states and controls in the nodes of the structure to be scalars, and therefore the elements of matrixes $\widetilde{\Omega}, \tilde{T}, \overline{\Omega}, \bar{T}$ are scalars. Generally, when states and controls are vectors, the elements of matrixes $\widetilde{\Omega}, \tilde{T}, \overline{\Omega}, \bar{T}$ will be matrixes.

*Notes.*

1. The difference of the system (1) from a similar system in the book [1] consists in that the $a_j$ node which participates in recording the $i$ equation of the system can generate the summand both in the left-hand and the right-hand side of the equation. The left-hand side summand contains a synthesizable (calculated) coefficient and corresponds to the soft connection between node $a_i$ and node $a_j$. The right-hand side summand contains the given coefficient and corresponds to the hard connection of node $a_i$ with node $a_j$. In [1] the $a_j$ node (in the equation for the $a_i$ node) could generate the summand only in one side of the equation.
2. The system (1) is written in the form which is convenient for further identification. After the identification, the synthesized and given coefficients with identical unknowns $x_i$ (or $u_i$) can be merged.
3. Generally speaking, the left-hand side of the system may contain an unknown vector $\tilde{C}$, and the right one - a given vector $\bar{C}$. But such a system can always be written as (1). For this purpose, it is necessary to formally introduce two additional nodes into the system which enter the neighborhoods of all nodes by states, while their own neighborhoods are empty. All the remaining nodes need to be considered soft-connected with the first formal node and hard -connected with the second one. At

the same time, the corresponding column of matrix $\widetilde{\Omega}$ (a column of calculated coefficients) is equal to $\tilde{C}$ and the corresponding column of matrix $\overline{\Omega}$ (a column of given coefficients) is equal to $\bar{C}$. All data by states of the formal nodes are unit $(x_{n+1}^k = x_{n+2}^k = 1)$ and all data by controls are zero $(v_{n+1}^k = v_{n+2}^k = 0)$. Such formal nodes do not generate the additional equations in the system.

In coordinate representation the linear symmetric neighborhood system takes the form

$$\sum_{q=1}^{n}[(\widetilde{\omega}_{iq}\tilde{s}_{iq})x_j + (\tilde{\tau}_{iq}\tilde{u}_{iq})v_q] = \sum_{q=1}^{n}[(\overline{\omega}_{iq}\bar{s}_{iq})x_q + (\bar{\tau}_{iq}\bar{u}_{iq})v_q] \qquad (2)$$

where $\widetilde{\omega}_{iq} = 0$ if $\tilde{s}_{iq} = 0$, $\tilde{\tau}_{iq} = 0$ if $\tilde{u}_{iq} = 0$, $\overline{\omega}_{iq} = 0$ if $\bar{s}_{iq} = 0$ and $\bar{\tau}_{iq} = 0$ if $\bar{u}_{iq} = 0$.

## THE SYNTHESIS (IDENTIFICATION) OF THE LINEAR NEIGHBORHOOD SYSTEM

Let us consider the task of finding coefficients $\widetilde{\omega}_{iq}$ and $\tilde{\tau}_{iq}$ of the neighborhood system (2) according to the experimental data, tuples $[X^k, V^k]$, $k = 1, \dots, K$. The substitution of data tuples in the equations of the neighborhood model (2) leads to the $nK$ system of linear equations for the unknown coefficients $\widetilde{\Omega}$ and $\tilde{T}$ of the neighborhood model. Generally speaking, in a neighborhood model there may be additional conditions connecting the desired coefficients of different equations of the model, e.g. the condition of the symmetry or antisymmetry of matrixes $\widetilde{\Omega}$ and $\tilde{T}$. If there are no such conditions, then the system of equations for finding the coefficients of the neighborhood model falls into $n$ systems, with one system per each node of the model:

$$\sum_{q=1}^{n}[\widetilde{\omega}_{iq}(\tilde{s}_{iq}x_j^k) + \tilde{\tau}_{iq}(\tilde{u}_{iq}v_q^k)] = b_i^k \qquad (3)$$

where $b_i^k = \sum_{q=1}^{n}[\overline{\omega}_{iq}(\bar{s}_{iq}x_q^k) + \bar{\tau}_{iq}(\bar{u}_{iq}v_q^k)]$. It should be stressed that the unknowns in the system (3) are the desired coefficients $\widetilde{\omega}_{iq}$ and $\tilde{\tau}_{iq}$ of the $i$ equation of the neighborhood model, while the numbers $\tilde{s}_{iq}x_j^k$ and $\tilde{u}_{iq}v_q^k$ are coefficients. The number of the (nonzero) unknowns $\widetilde{\omega}_{iq}$ and $\tilde{\tau}_{iq}$ in the system (3) will be denoted by $r_i$. In a non-degenerate case, i.e. with the maximum rank of the system matrix, the coefficients $\widetilde{\omega}_{iq}$ and $\tilde{\tau}_{iq}$ $i$ may be found as:

a) The normal solution of the underdetermined system (the solution with the minimum norm), if $K < r_i$;
b) The (unique) solution of the system in case of $K = r_i$;
c) The pseudo-solution (last square solution) of the overdetermined system, if $K > r_i$.

If the rank of the system matrix is not maximal, then the desired coefficients can be found as the pseudo-solution of the system: the norm-minimal vector minimizing the residual norm.

## METRIC ON A SET OF NODES OF THE NEIGHBORHOOD STRUCTURE

For each node $a_i$of the neighborhood structure we set $D_i = C_i/\|C_i\|$, where

$$C_i = [\tilde{s}_{i1}, \dots, \tilde{s}_{in}, \bar{s}_{i1}, \dots, \bar{s}_{in}, \tilde{u}_{i1}, \dots, \tilde{u}_{in}, \bar{u}_{i1}, \dots, \bar{u}_{in}]. \quad (4)$$

The metric will be defined on a set of nodes of the neighborhood structure by the formula

$$r_{ij} = \sqrt{\|D_i - D_j\|^2 + \|X_i - X_j\|^2 + \|V_i - V_j\|^2}. \quad (5)$$

The first summand under the root evaluates the similarity of node neighborhoods *taking into account weighed connections*. The node entering the neighborhoods by states (controls) of nodes $a_i$ and $a_j$ or not entering any of these neighborhoods, will have a zero contribution to $\|D_i - D_j\|$. The $a_q$ node entering only one of these neighborhoods will have a contribution proportional to the weight of the corresponding connection. The sum of the two following summands under the root is the square of the distance between the vectors of the given nodes $a_i$ and $a_j$. Thus, the defined metric calculates the proximity of nodes according to observation data and to their weighed connections (neighborhoods) in the neighborhood system.

*Note*. Another approach to defining the measure of proximity of the neighborhood structure nodes according to connections and data was discussed in [2]. It should be noted that the measure of $d_{ij}$ defined in [2] is not, generally speaking, a metric (the triangle inequality is not met).

### NEIGHBORHOOD FACTOR STRUCTURES

Using the $r_{ij}$ metric defined above, it is possible to break the nodes of a neighborhood structure into clusters by arithmetic means of a clustering algorithm. In any case, further it is necessary to recalculate the adjacency matrix of the neighborhood structure and the data matrix. The mean of the data vectors of all the nodes comprising a cluster can be taken as the data vector (by control or state). Further, the neighborhood of the cluster *I* by states (controls) will be considered as one consisting of all clusters $J$, so that at least one of the nodes of the cluster $J$ enters the neighborhood of at least one of the nodes of the cluster $I$. The weight of the corresponding soft connection of the cluster *I* with the cluster $J$ will be considered to be equal to the arithmetic mean of the weights of all soft connections of the nodes from the cluster *I* with the nodes from the cluster $J$. Similarly, the weight of the corresponding hard connection of the cluster *I* with the cluster $J$ will be considered to be equal to the arithmetic mean of the weights of all hard connections of the nodes from the cluster *I* with the nodes from the cluster $J$. The given coefficients of hard connections of clusters can be defined as arithmetic means of the corresponding given coefficients of hard connections of the nodes from $I$ with the nodes from $J$.

Acknowledgments: The work is supported by the Russian Fund for Basic Research (project 16-07-00854 a).

## REFERENCES

[1] Blyumin S.L., Shmyrin A.M. Neighborhood system. Lipetsk: LEGI, 2005.

[2] Shmyrin, A.M., Kosareva, A.S. The measure of similarity in solving the problem of clustering neighborhood structures // Modern informatization problems in the technological and telecommunication systems analysis and synthesis: Proceedings of the XXI-th International Open Science Conference (Yelm, WA, USA, Januar 2016). 2016. – p.341-346.