CONSTRUCTION OF HIERARCHICAL CLASSIFICATION BY SIMILARITY MATRIX

G. Tsitsiashvili, M. Osipova

Institute for Applied Mathematics Far Eastern Branch of RAS, Vladivostok, Russia
Far Eastern Federal University, Vladivostok, Russia

e-mails: guram@iam.dvo.ru, mao1975@list.ru

ABSTRACT

In this paper a problem of hierarchical classification of some objects by similarity matrix solution is solved. This approach gives single solution of classification problem. Each hierarchical level is defined by some critical value of a similarity. Using critical value the similarity matrix is transformed into contiguity matrix of some no oriented graph in which connectivity components are constructed. Increasing successfully critical values it is possible to define hierarchical classification of initial objects. This approach is closely connected with reliability theory and mathematical statistics in which a reaching of critical value is one of important problems.

1 INSTRUCTIONS

A problem of a classification by a matrix of pair similarities (differences) between objects is widely used in data processing (Aivazian, Enukov, Meshalkin, 1983). These problems usually are considered as a solution of some minimum – maximum problem in which a similarity between objects in a class is maximized (difference between objects in a class are minimized) and a similarity between classes is minimized (difference between classes is maximized). One of disadvantages of such a statement of a question is no uniqueness of this problem solution and difficulties with an enumeration of all possible solutions.

But last years in different applications a tendency to solve classification problem in a connection with a construction of hierarchical classification intensifies. Such statement of a question increases demands to a uniqueness of classification problem. In this paper the uniqueness is reached by a transformation of similarity matrix into zero – one matrix via a comparison of similarity matrix elements with some critical value. This zero – one matrix becomes contiguity matrix of some no oriented graph. Then connectivity components of this graph are constructed. These components are identified with some objects classes. To define hierarchical classification of the objects critical values are increased. As a result these classes are divided into subclasses and so on. In such a way hierarchical classification is constructed by matrix of pair similarities (of pair differences).

Another applied problem of the classification is a definition of upper boundary "supremum" for similarity matrix (of lower boundary "infinum" for difference matrix) of critical values for which classification procedure gives single solution. This problem may be solved using the method of dichotomous division. For this aim on initial step we take a pair of critical values: zero and a maximum of pair similarities (a minimum and a maximum of pair differences).

Consequently hierarchical classification transforms into a definition of connectivity components in some no oriented graph. Known algorithms (Kormen, Leizerson, Rivest, 2004), (Graham, Hell, 1985) of connectivity components construction are based on a search into a depth and into a width. A disadvantage of these algorithms is in repeated calls to earlier considered graph edges in "search tree" and a definition of for the nodes all nodes contiguity with them. In this paper we suggest algorithm which has not these disadvantages. If considered graph is connected this algorithm is similar with algorithm of spanning tree construction (Eppstein, 1999).

2 ALGORITHM OF HIERARCHICAL CLASSIFICATION

Consider n objects and denote μ_{ij} their similarity measure i, j, $i \neq j.$ Then similarity matrix between these objects is $M = \| \mu_{ij} \|_{i,j=1}^{n}$. It consists of nonnegative numbers with $\mu_{ii} = \overline{m},$ where $\overline{m} = \max\limits_{1 \le i \neq j \le n} \mu_{ij} + 1.$ To each integer number $k, 0 \le k \le \overline{m},$ contrast matrix $M^{(k)} = \| m_{ij}^{(k)} \|_{i,j=1}^{n},$ where $m_{ij}^{(k)} = 1,$ if $\mu_{ij} \ge k,$ else $m_{ij}^{(k)} = 0.$ The matrix $M^{(k)}$ consists of zeros and units and may be considered as contiguity matrix of some no oriented graph $G^{(k)}$ with n nodes which designate initial objects.

In the graph $G^{(k)}$ construct connectivity components $J_1^{(k)}, J_2^{(k)}, \ldots, J_{n(k)}^{(k)}$, so that for any two nodes $i, j \in J_t^{(k)}$ in the graph $G^{(k)}$ there is a way which connect them. If $i \in J_t^{(k)}, j \in J_l^{(k)}, t \neq l$ then there are not ways which connect the nodes i, j in the graph $G^{(k)}$. Remark that when we transit from k to $k+1$ each set $J_i^{(k+1)}$ completely contains to some set $J_j^{(k)}$ or does not intersect with it.

Consequently the subsets $J_1^{(k)}, J_2^{(k)}, \ldots, J_{n(k)}^{(k)}$ create a decomposition of the set $\{1, \ldots, n\}$ into classes by the levels $k, 0 \le k \le \overline{m},$

$$\{J_1^{(0)}, J_2^{(0)}, \ldots, J_{n(0)}^{(0)}\},$$
$$\{J_1^{(1)}, J_2^{(1)}, \ldots, J_{n(1)}^{(1)}\}, \ldots,$$
$$\{J_1^{(\overline{m})}, J_2^{(\overline{m})}, \ldots, J_{n(\overline{m})}^{(\overline{m})}\},$$

in which for any class $J_t^{(k+1)}$ of the level $k+1$ there is the class $J_l^{(k)},$ satisfying the inclusion $J_t^{(k+1)} \subseteq J_l^{(k)}.$ Further construct the tree D with the hight $\overline{m},$ its root is the node $J_1^{(0)} = \{1, \ldots, n\}.$ On the level $k = 1$ consider the nodes $J_1^{(1)}, J_2^{(1)}, \ldots, J_{n(1)}^{(1)}$ and connect them by edges with the node $J_1^{(0)}$. On the level $k = 2$ consider nodes $J_1^{(2)}, J_2^{(2)}, \ldots, J_{n(2)}^{(2)}$ and connet the node $J_t^{(2)}$ by the edge with the node $J_l^{(1)}$ of the level $1,$ if there is the inclusion $J_t^{(2)} \subseteq J_l^{(1)}.$ This procedure continues to the level $\overline{m},$ on which the set $\{1, \ldots, n\}$ is divided into n one node subsets. To simplify the description of the tree it is possible to replace the inclusion $J_t^{(k+1)} \subseteq J_l^{(k)}$ by the inclusion $J_t^{(k+1)} \subset J_l^{(k)}.$ If $J_t^{(k+1)} = J_l^{(k)}$ then the nodes $J_t^{(k+1)}, J_l^{(k)}$ of the tree D are glued.

To construct connectivity components in no oriented graph g with n nodes we use the following algorithm. On the step 1 take the node 1 and construct the connectivity component $K_1^{(1)} = \{1\}$. Assume that on the step $t-1$ the set of nodes $\{1, \ldots, t-1\}$ is divided into connectivity components $K_i^{(t-1)}, i \in L_{t-1}$:

$$K_i^{(t-1)} \bigcap K_j^{(t-1)} = \varnothing, i \neq j, \;\; i, j \in J,$$
$$\bigcup_{i \in L_{t-1}} K_i^{(t-1)} = \{1,...,t-1\}.$$

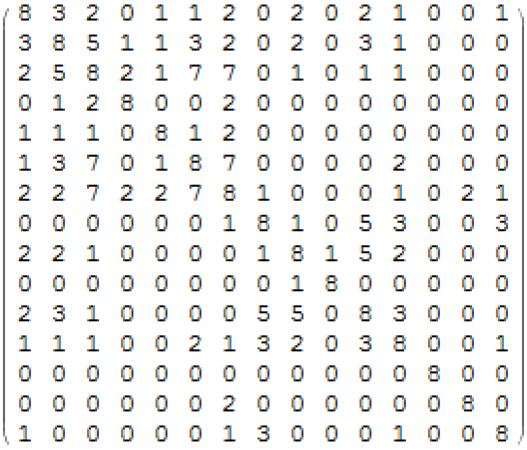
On the step t consider the next node t and calculate $c_i = \max_{j \in K_i^{(t-1)}} m_{tj}^{(k)}, i \in L_{t-1}$ and put

$$I = \{i \in L_{t-1} : c_i = 1\}, \; K_i^{(t)} := K_i^{(t-1)}, i \in L_{t-1} / I,$$
$$K_t^{(t)} := \{t\} \bigcup \left[\bigcup_{i \in I} K_i^{(t)} \right], \; L_t := (L_{t-1} / I) \bigcup \{t\}.$$

This means that the classes $K_i^{(t-1)}, i \in L_{t-1},$ with which the node t is connected by some edges, are aggregated with t into new class $K_t^{(t)}$.

3 NUMERICAL EXAMPLE

Assume that similarity matrix of 15 objects has the form ($\overline{m} = 8$):



$$\begin{pmatrix}
8 & 3 & 2 & 0 & 1 & 1 & 2 & 0 & 2 & 0 & 2 & 1 & 0 & 0 & 1 \\
3 & 8 & 5 & 1 & 1 & 3 & 2 & 0 & 2 & 0 & 3 & 1 & 0 & 0 & 0 \\
2 & 5 & 8 & 2 & 1 & 7 & 7 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 2 & 8 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 8 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 3 & 7 & 0 & 1 & 8 & 7 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
2 & 2 & 7 & 2 & 2 & 7 & 8 & 1 & 0 & 0 & 0 & 1 & 0 & 2 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 8 & 1 & 0 & 5 & 3 & 0 & 0 & 3 \\
2 & 2 & 1 & 0 & 0 & 0 & 0 & 1 & 8 & 1 & 5 & 2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 8 & 0 & 0 & 0 & 0 & 0 \\
2 & 3 & 1 & 0 & 0 & 0 & 0 & 5 & 5 & 0 & 8 & 3 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 2 & 1 & 3 & 2 & 0 & 3 & 8 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 3 & 0 & 0 & 0 & 1 & 0 & 0 & 8
\end{pmatrix}$$

For any critical level $k = 0, 1, ..., 8$ the graph $G^{(k)}$ has the following connectivity components (connectivity components with single element do not repeat on successive levels):
the level $k = 0$: $J_1^{(0)} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15\}$
the level $k = 1$: $J_1^{(1)} = \{1,2,3,4,5,6,7,8,9,10,11,12,14,15\}$, $J_2^{(1)} = \{13\}$
the level $k = 2$: $J_1^{(2)} = \{1,2,3,4,5,6,7,8,9,11,12,14,15\}$, $J_2^{(2)} = \{10\}$
the level $k = 3$: $J_1^{(3)} = \{1,2,3,6,7,8,9,11,12,15\}$, $J_2^{(3)} = \{4\}$, $J_3^{(3)} = \{5\}$, $J_4^{(3)} = \{14\}$
the level $k = 4$: $J_1^{(4)} = \{1\}$, $J_2^{(4)} = \{2,3,6,7\}$, $J_3^{(4)} = \{8,9,11\}$, $J_4^{(4)} = \{12\}$, $J_5^{(4)} = \{15\}$
the level $k = 5$: $J_1^{(5)} = \{2,3,6,7\}$, $J_2^{(5)} = \{8,9,11\}$
the level $k = 6$: $J_1^{(6)} = \{2\}$, $J_2^{(6)} = \{3,6,7\}$, $J_3^{(6)} = \{8\}$, $J_4^{(6)} = \{9\}$, $J_5^{(6)} = \{11\}$
the level $k = 7$: $J_1^{(7)} = \{3,6,7\}$
the level $k = 8$: $J_1^{(8)} = \{3\}$, $J_2^{(8)} = \{6\}$, $J_3^{(8)} = \{7\}$
Then we construct the tree D with the hight 7 with glued nodes: $J_2^{(6)}$ with $J_1^{(7)}$, $J_2^{(4)}$ with $J_1^{(5)}$, $J_3^{(4)}$ with $J_2^{(5)}$ (Fig. 1).

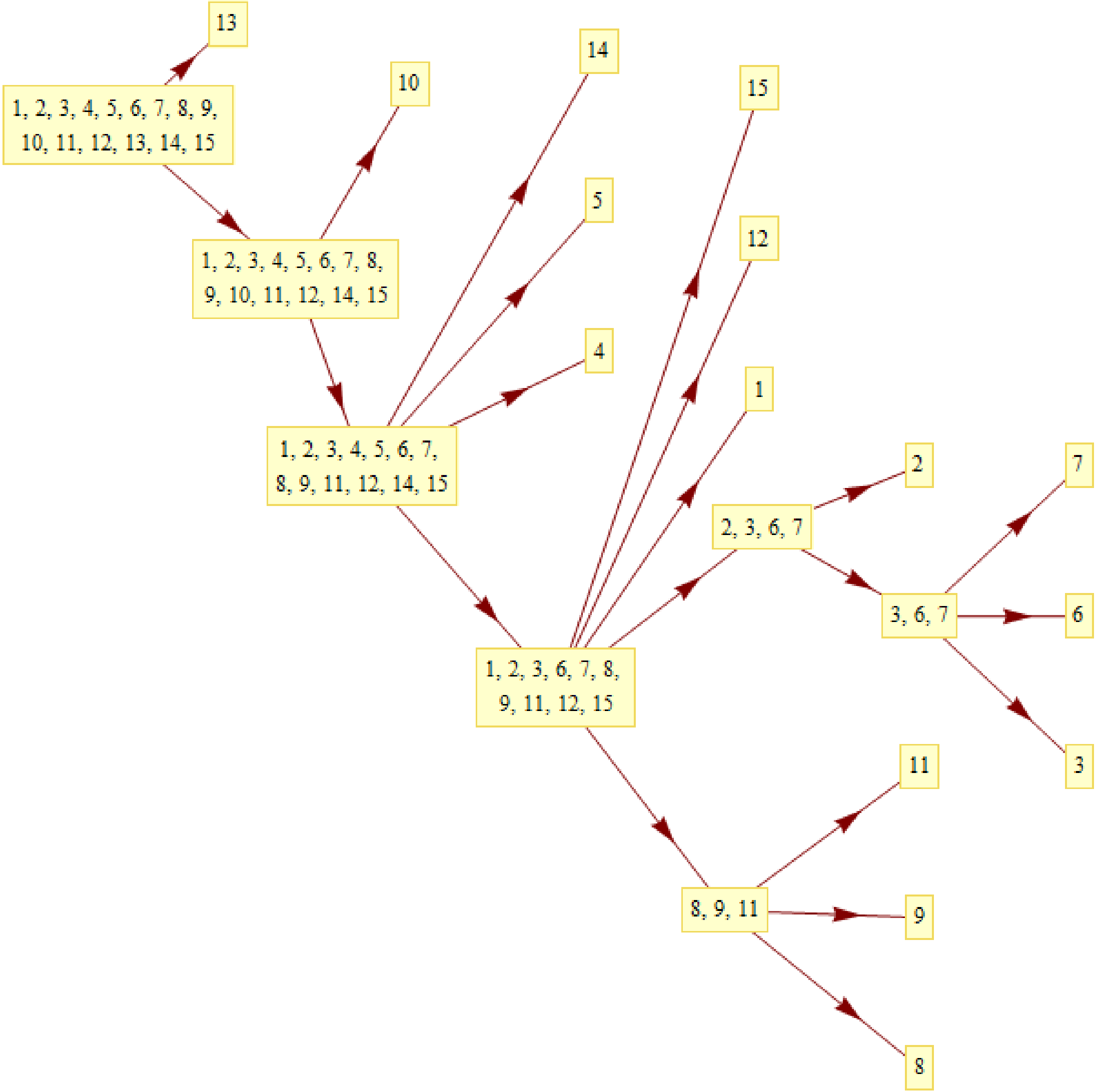


Figure 1. The tree D with the hight 7.

4 REFERENCES

1. Aivazian S.A., Enukov I.S., Meshalkin L.D. 1983. *Applied statistics. Bases of modeling and initial data processing*. Moscow: Finances and statistics. (In Russian).

2. Eppstein D. Spanning trees and spanners. In Sack J.R. and Urrutia J. 1999. *Handbook of Computational Geometry*. Elsevier. P. 425-461.

3. Graham R.L., Hell P. 1985. On the history of the minimum spanning tree problem. *Annals of the History of Computing* 7 (1): 43-57.

4. Kormen T., Leizerson Ch., Rivest R. 2004. *Algorithms: construction and analysis*. Moscow: Laboratory of basic knowledges. (In Russian).