15th Int. Workshop on Generalized Nets
Burgas, 16 October 2014, 39–48

Hierarchical generalized net model of the process of selecting a method for clustering

Veselina Bureva[1], Evdokia Sotirova[1], Krassimir Atanassov[2]

[1] "Prof. Asen Zlatarov" University
1 "Prof. Yakimov" Blvd., Burgas – 8010, Bulgaria
e-mails: esotirova@btu.bg, vbureva@btu.bg

[2] Department of Bioinformatics and Mathematical Modelling
Institute of Biophysics and Biomedical Engineering
Bulgarian Academy of Sciences
e-mail: krat@bas.bg

Abstract: In the current paper is presented a generalized net model of selecting a method for clustering. That GN-model is a subnet of the transition Z_6 of the GN-model "Hierarchical generalized net model of the process of clustering". The generalized net that is presented in the article can be useful in analizing, managing and optimizing the process of clustering.
Keywords: Clustering, Generalized Net, Data Mining, Knowledge Discovery, Segmentation
AMS Classification: 68Q85, 62H30.

1 Introduction

Let us suppose that is given a dataset with n points in a d-dimensional space and it is given the number of desired clusters k. The task of clustering is to partitioning the dataset in k-groups. The number of clusters isn't known at any time. It depends of the type of clustering, of the selected method and the algorithm which is used. In the present paper are analyzed the main methods for clustering. The most popular of them are hierarchical clustering, partitioning clustering, density-based clustering, grid-based clustering, model-based clustering. Each of them has own methods and algorithms that are used to group clusters. The whole process of clustering contains tree summarized steps – training, validation and testing. The input data is separated in three parts - training set, validation set and testing set. Previously the training set is preprocessed which means that it will not contain missing values, outliers (noise). The training set, the clustering method and the criteria are used to form clusters. The validation set and the criteria are used to set the received clusters. The result of clustering is applied to testing set to predict future values. Ordinarily the criteria are similarity and distance-based measures, number for clusters [1, 4, 11].

2 Generalized net model

The concept for Generalized nets is introduced in [2, 3]. There are several methods and algorithms for knowledge discovery that are already modeled by generalized nets [5–10]. In this paper is constructed a generalized net model of the process of selecting a method for clustering. The GN is presented on the Fig. 1. It contains 7 transitions and 41 places.

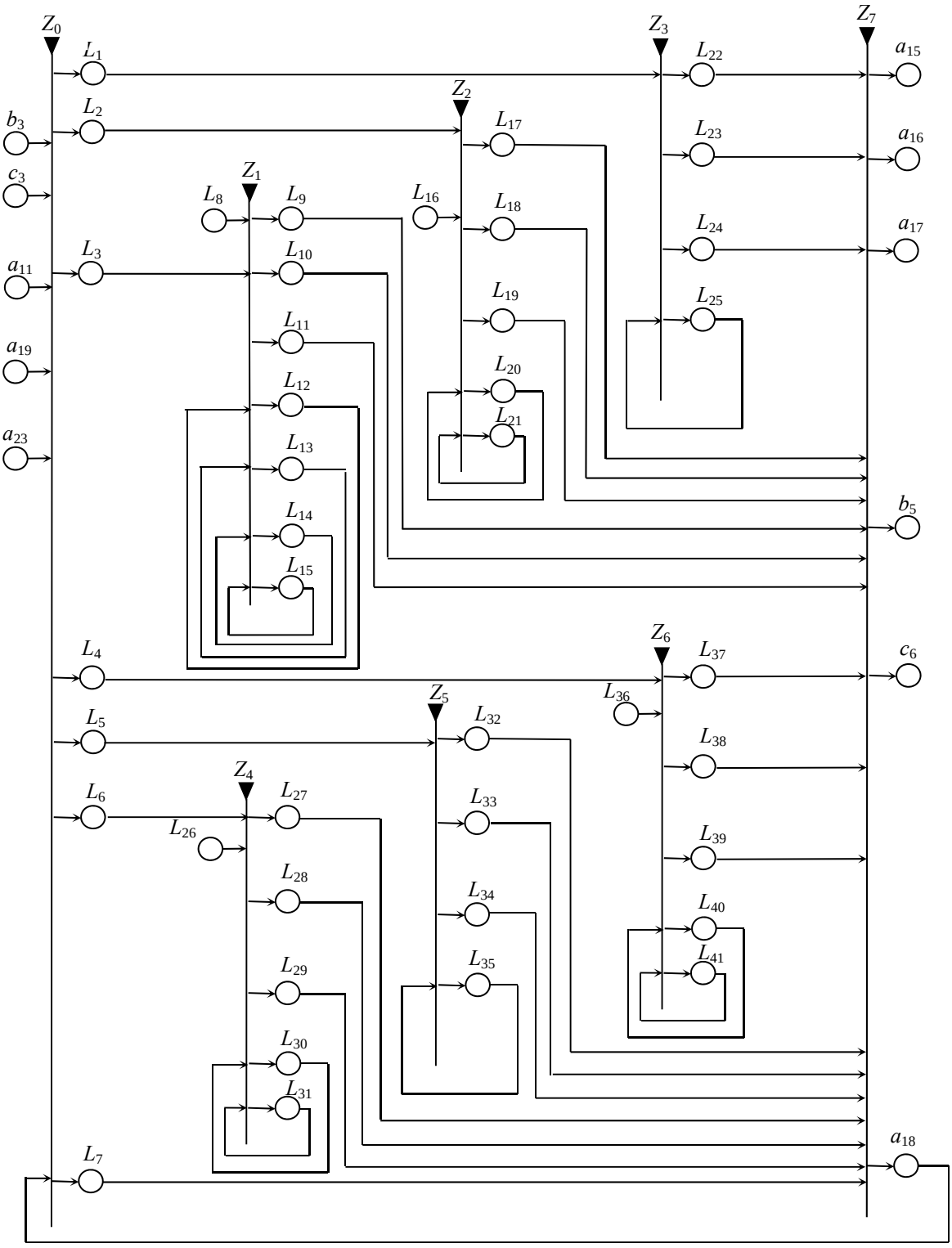

Figure 1. Hierarchical generalized net model of the process of selecting a method for clustering

It can replace the transition Z_6 in a generalized net of the process of clustering using operator H_3. The set of transitions A is the following:

$$A_1 = \{Z_0, Z_1, Z_2 , Z_3 , Z_4 , Z_5 , Z_6 , Z_7\},$$

where the transitions describe the processes:

- Z_0 - "input parameters";
- Z_1 - "hierarchical methods for clustering";
- Z_2 - "partitioning methods for clustering";
- Z_3 - "density-based clustering";
- Z_4 - "grid-based clustering";
- Z_5 - " model-based clustering";
- Z_6 - " other methods for clustering - fuzzy clustering, soft computing";
- Z_7 - "output data".
-

The transition Z_0 has the following form:

$$Z_0 = \langle\{b_3, c_3, a_{11}, a_{19}, a_{23}, L_{25}\}, \{L_1, L_2, L_L\text{-}3, L_4, L_5, L_6, L_7\}, R_0,$$
$$\vee(\wedge(b_3, c_3, a_{11}), a_{19}, a_{23}, L_{25}) \rangle,$$

where:

$$R_0 = \begin{array}{c|ccccccc} & L_1 & L_2 & L_3 & L_4 & L_5 & L_6 & L_7 \\ \hline b_3 & W_{3,1} & W_{3,2} & W_{3,3} & W_{3,4} & W_{3,5} & W_{3,6} & false \\ c_3 & W_{3,1} & W_{3,2} & W_{3,3} & W_{3,4} & W_{3,5} & W_{3,6} & false \\ a_{11} & W_{11,1} & W_{11,2} & W_{11,3} & W_{11,4} & W_{11,5} & W_{11,6} & false \\ a_{19} & W_{19,1} & W_{19,2} & W_{19,3} & W_{19,4} & W_{19,5} & W_{19,6} & false \\ a_{23} & W_{23,1} & W_{23,2} & W_{23,3} & W_{23,4} & W_{23,5} & W_{23,6} & false \\ L_{25} & W_{25,1} & W_{25,2} & W_{25,3} & W_{25,4} & W_{25,5} & W_{25,6} & false \end{array},$$

and:

i = {1,2,...,6}, where:
i = 1-"hierarchical clustering";
i = 2-"partitioning clustering";
i = 3-"density-based clustering";
i = 4-"grid-based clustering";
i = 5-"model-based clustering";
i = 6-"other methods for clustering";

The predicates in the index matrix R_0 have the following meaning:

- $W_{3,1} = W_{3,2} = W_{3,3} = W_{3,4} = W_{3,5} = W_{3,6}$ = "selected criteria for clustering of type i" (from place b_3);
- $W_{3,1} = W_{3,2} = W_{3,3} = W_{3,4} = W_{3,5} = W_{3,6}$ = = "selected method and algorithm for clustering of type i" (from place c_3) ;
- $W_{11,1} = W_{11,2} = W_{11,3} = W_{11,4} = W_{11,5} = W_{11,6}$ = " selected data for clustering of type i";
- $W_{19,1} = W_{19,2} = W_{19,3} = W_{19,4} = W_{19,5} = W_{19,6}$ = "there are data for validation clustering of type i";

- $W_{23,1} = W_{23,2} = W_{23,3} = W_{23,4} = W_{23,5} = W_{23,6}$ = " there are data for testing clustering of type i";
- $W_{25,1} = W_{25,2} = W_{25,3} = W_{25,4} = W_{25,5} = W_{25,6} = W_{25,7}$ = "it is necessary new data, criteria or method";

The tokens from places $b_3, c_3, a_{11}, a_{19}, a_{23}, L_{25}$ enter in transition Z_0. In places L_1, L_2, $L_{L\text{-}3}$, L_4, L_5, L_6, L_7 enter tokens with characteristics:

- "*data, criteria and method for hierarchical clustering*" in place L_1,
- "*data, criteria and method for partitioning clustering*" in place L_2,
- "*data, criteria and method for density-based clustering*" in place L_3,
- "*data, criteria and method for grid-based clustering*" in place L_4,
- "*data, criteria and method for model-based clustering*" in place L_5,
- "*data, criteria and method for other methods for clustering*" in place L_6,
- "*data, criteria and method for other (second) types of clustering*" in place L_7.

The σ-token entering in the net via place L_8 with initial characteristic: "*method for creating clusters - 'bottom-top' or 'top-bottom'* ".

The transition Z_1 has the following form:

$$Z_1 = \langle \{ L_8, L_3, L_{12}, L_{13}, L_{14}, L_{15}\}, \{L_9, L_{10}, L_{11}, L_{12}, L_{13}, L_{14}, L_{15}\}, R_1, \vee(\wedge(L_8, L_3), L_{12}, L_{13}, L_{14}, L_{15}) \rangle,$$

where:

$R_1 =$

	L_9	L_{10}	L_{11}	L_{12}	L_{13}	L_{14}	L_{15}
L_8	*false*	*false*	*false*	*false*	*false*	*false*	*true*
L_3	*false*	*false*	*false*	*false*	*false*	*false*	*true*
L_{12}	$W_{12,9}$	$W_{12,10}$	$W_{12,11}$	$W_{12,12}$	*false*	*false*	*false*,
L_{13}	*false*	*false*	*false*	$W_{13,12}$	$W_{13,13}$	*false*	*false*
L_{14}	*false*	*false*	*false*	*false*	$W_{14,13}$	$W_{14,14}$	*false*
L_{15}	*false*	*false*	*false*	*false*	*false*	$W_{15,14}$	$W_{15,15}$

and:

- $W_{12,9}$ = " κ-clusters from hierarchical clustering are received ";
- $W_{12,10}$ = " n-points between the clusters from hierarchical clustering are received ";
- $W_{12,11}$ = "it is necessary other type of clustering ";
- $W_{12,12} = \neg(W_{12,9} \wedge W_{12,10} \wedge W_{12,11})$;
- $W_{13,12}$ = "the type of hierarchical clustering is determined - 'agglomerative' or 'divisive'";
- $W_{13,13} = \neg W_{13,12}$;
- $W_{14,13}$ = " the way to calculate the similarity measures is determined - 'single-linkage', 'complete linkage', 'group average '";
- $W_{14,14} = \neg W_{14,13}$;
- $W_{15,14}$ = " the direction of clustering is determined -'bottom-top', 'top-bottom'";
- $W_{15,15} = \neg W_{15,14}$.

At the first activation of transition the token from place L_8 entering in place L_{15} don't obtain new characteristic. In the next time moment the token from place L_{15} generates new one that enters in place L_{14} with characteristic: *"direction of clustering - 'bottom-top', 'top-bottom'"*. At the third activation of transition the token from place L_{14} generates new one that enters in place L_{13} with characteristic: "*the way to calculate the similarity measures*". In the next time moment the token from place L_{13} generates new one that enters in place L_{12} with characteristic:*"type of clustering* - 'agglomerative' or 'divisive'*"*. In the next activation the token from place L_{12} generates three new tokens that enter in places L_9, L_{10} and L_{11} with characteristics: *"received clusters from hierarchical clustering"* in place L_9, *"data fallen between the clusters in hierarchical clustering "* in place L_{10}, *"output data from hierarchical clustering to choosing a new clustering method" in place* L_{11}.

The σ-token entering in place L_5 with initial characteristic: *"given number of clusters "*.
The transition Z_2 has the form:

$$Z_2 = \langle \{ L_{16}, L_2, L_{20}, L_{21}\}, \{L_{17}, L_{18}, L_{19}, L_{20}, L_{21}\}, R_2, \vee(L_{16}, L_2, L_{20}, L_{21}) \rangle,$$

where:

$$R_2 = \begin{array}{c|ccccc} & L_{17} & L_{18} & L_{19} & L_{20} & L_{21} \\ \hline L_{16} & false & false & false & true & false \\ L_2 & false & false & false & false & true \\ L_{20} & W_{20,17} & W_{20,18} & W_{20,19} & W_{20,20} & false \\ L_{21} & false & false & false & W_{21,20} & W_{21,21} \end{array},$$

and:

- $W_{20,17}$ = " κ-clusters from partitioning clustering are received ";
- $W_{20,18}$ = " n-points between the clusters from partitioning clustering are received ";
- $W_{20,19}$ = " it is necessary other type of clustering ";
- $W_{20,20} = \neg(W_{20,17} \wedge W_{20,18} \wedge W_{20,19})$;
- $W_{21,20}$ = "there is selected number for clusters".
- $W_{21,21} = \neg W_{21,20}$

At the first activation of the transition the token entering in place L_{21} (from place L_{16}) don't obtain new characteristic. At the second activation the token from place L_{21} generates new one that enters in place L_{20} with characteristic: *"number of clusters"*. In this time moment the token from place L_2 enters in place L_{20} and generate three new tokens. They enter in places L_{17}, L_{18} and L_{19} with characteristics: *"received clusters from partitioning clustering"* in place L_{17}, *"received data between clusters from partitioning clustering"* in place L_{18}, *"output data from partitioning clustering for choosing other clustering method"* in place L_{19}.

The transition Z_3 has the form:

$$Z_3 = \langle \{ L_1, L_{25}\}, \{L_{22}, L_{23}, L_{24}, L_{25}\}, R_3, \vee(L_1, L_{25}) \rangle,$$

where:

$$R_3 = \begin{array}{c|cccc} & L_{22} & L_{23} & L_{24} & L_{25} \\ \hline L_1 & false & false & false & true \\ L_{25} & W_{25,22} & W_{25,23} & W_{25,24} & W_{25,25} \end{array},$$

and:

- $W_{25,22}$ = " κ - clusters from density-based clustering are received ";
- $W_{25,23}$ = " n - points between the clusters from density-based clustering are received ";
- $W_{25,24}$ = " it is necessary other type of clustering ";
- $W_{25,25} = \neg(W_{25,22} \wedge W_{25,23} \wedge W_{25,24})$;

The token entering in place L_{25} (from place L_1) don't obtain new characteristic. The token from place L_{25} generates three new token that enter in places L_{22}, L_{23} and L_{24} with characteristics: *"received clusters from density-based clustering"* in place L_{22}, *"received data between clusters from density-based clustering "* in place L_{23}, *"output data from density-based clustering for choosing other clustering method"* in place L_{24}.

The σ-token entering in the net via place L_{26} with initial characteristic:*"dimension".*

The transition Z_4 has the form:

$$Z_4 = \langle \{L_{26}, L_6, L_{30}, L_{31}\}, \{L_{27}, L_{28}, L_{29}, L_{30}, L_{31}\}, R_4, \vee(\wedge(L_{26}, L_6), L_{30}, L_{31}) \rangle,$$

where:

$$R_4 = \begin{array}{c|ccccc} & L_{27} & L_{28} & L_{29} & L_{30} & L_{31} \\ \hline L_{26} & false & false & false & false & true \\ L_6 & false & false & false & true & false \\ L_{30} & W_{30,27} & W_{30,28} & W_{30,29} & W_{30,30} & false \\ L_{31} & false & false & false & W_{31,30} & W_{31,31} \end{array},$$

and:

- $W_{30,27}$ = " κ - clusters from grid-based clustering are received ";
- $W_{30,28}$ = " n - points between the clusters from grid-based clustering are received ";
- $W_{30,29}$ = " it is necessary other type of clustering ";
- $W_{30,30} = \neg(W_{30,27} \wedge W_{30,28} \wedge W_{30,29} \wedge W_{30,29})$.
- $W_{31,30}$ = "there is a selected number for clusters";
- $W_{31,31} = \neg W_{31,30}$.

At the first activation of the transition the token from place L_{26} enters in place L_{31} and don't obtain new characteristic. At the second activation the token from place L_{31} generates new token that enters in place L_{30} with characteristic: *"selected dimension for mapping the data".*

At the third activation the token from place L_{30} generates three new tokens that enters in places L_{27}, L_{28} and L_{29} with characteristics: *"received clusters from grid-based clustering "* in place L_{27}, *" received data between clusters from grid-based clustering "* in place L_{28}, *"output data from grid-based clustering for choosing other clustering method"* in place L_{29}.

The transition Z_5 has the form:

$$Z_5 = \langle \{ L_5, L_{35} \}, \{ L_{32}, L_{33}, L_{34}, L_{35} \}, R_5, \vee(L_5, L_{36}) \rangle,$$

where:

$$R_5 = \begin{array}{c|cccc} & L_{32} & L_{33} & L_{34} & L_{35} \\ \hline L_5 & false & false & false & true \\ L_{35} & W_{35,32} & W_{35,33} & W_{35,34} & W_{35,35} \end{array},$$

and:

- $W_{35,32}$ = " κ - clusters from model-based clustering are received";
- $W_{35,33}$ = " n - points between the clusters from grid-based clustering are received ";
- $W_{35,34}$ = " it is necessary other type of clustering ";
- $W_{35,35} = \neg(W_{35,32} \wedge W_{35,33} \wedge W_{35,34})$.

At the first activation of the transition the token from place enters in place L_{35} and don't obtain new characteristic. At the second activation of the transition the token from place L_{35} generates three new tokens that enter in places L_{32}, L_{33} and L_{34} with characteristics: *"received clusters from model-based clustering "* in place L_{32}, *"received data between clusters from model-based clustering "* in place L_{33}, *"output data from model-based clustering for choosing other clustering method"* in place L_{34}.

The σ-token enters the net via place L_{36} with initial characteristic: *"other parameters".*

The transition Z_6 has the form:

$$Z_6 = \langle \{ L_{36}, L_4, L_{40}, L_{41} \}, \{ L_{37}, L_{38}, L_{39}, L_{40}, L_{41} \}, R_6, \vee(\wedge(L_{36}, L_4), L_{40}, L_{41}) \rangle,$$

where:

$$R_6 = \begin{array}{c|ccccc} & L_{37} & L_{38} & L_{39} & L_{40} & L_{41} \\ \hline L_{36} & false & false & false & false & true \\ L_4 & false & false & false & true & false \\ L_{40} & W_{40,37} & W_{40,38} & W_{40,39} & W_{40,40} & false \\ L_{41} & false & false & false & W_{41,40} & W_{41,41} \end{array},$$

and:

$W_{40,37}$ = " κ - clusters from clustering of other method are received ";

$W_{40,38}$ = " n - points between the clusters from clustering of other method are received ";

$W_{40,39}$ = " it is necessary other type of clustering ";

$W_{40,40} = \neg(W_{40,37} \wedge W_{40,38} \wedge W_{40,39})$

$W_{41,40}$ = "there are selected parameters";

$W_{41,41} = \neg W_{41,40}$

At the first activation of the transition the token from place L_{38} entering in place L_{41} and don't obtain new characteristic. At the second activation of the transition the token from place L_{41} generates new one with characteristic: "*selected other parameters*".

The token entering in place L_{41} (from place L_6) don't obtain new characteristic. At the third activation of the transition the token from place L_{41} generates three new tokens that enters in

places L_{38}, L_{39} and L_{40} with characteristics: *"received clusters from other method for clustering "* in place L_{38}, *"received data between clusters from other method for clustering "* in place L_{39}, *"*it is necessary other type of clustering *"* in place L_{40}.

The transition Z_7 has the form:

$$Z_7 = <\{ L_{22}, L_{23}, L_{24}, L_{17}, L_{18}, L_{19}, L_9, L_{10}, L_{11}, L_{37}, L_{38}, L_{39}, L_{32}, L_{33}, L_{34}, L_{27}, L_{28}, L_{29}, L_7\},$$
$$\{a_{15}, a_{16}, a_{17}, b_5, c_6, a_{18}\}, R_7,$$
$$\vee(L_{22}, L_{23}, L_{24}, L_{17}, L_{18}, L_{19}, L_9, L_{10}, L_{11}, L_{37}, L_{38}, L_{39}, L_{32}, L_{33}, L_{34}, L_{27}, L_{28}, L_{29}, L_7)>,$$

where:

$$R_{R-7} = \begin{array}{c|cccccc} & a_{15} & a_{16} & a_{17} & b_5 & c_6 & a_{18} \\ \hline L_{22} & W_{22,15} & false & false & false & false & false \\ L_{23} & false & W_{23,16} & false & false & false & false \\ L_{24} & false & false & false & false & false & W_{24,18} \\ L_{17} & W_{17,15} & false & false & false & false & false \\ L_{18} & false & W_{18,16} & false & false & false & false \\ L_{19} & false & false & false & false & false & W_{19,18} \\ L_9 & W_{9,15} & false & false & false & false & false \\ L_{10} & false & W_{10,16} & false & false & false & false \\ L_{11} & false & false & false & false & false & W_{11,18} \\ L_{37} & W_{37,15} & false & false & false & false & false \\ L_{38} & false & W_{38,16} & false & false & false & false \\ L_{39} & false & false & false & false & false & W_{39,18} \\ L_{32} & W_{32,15} & false & false & false & false & false \\ L_{33} & false & W_{33,16} & false & false & false & false \\ L_{34} & false & false & false & false & false & W_{34,18} \\ L_{27} & W_{27,15} & false & false & false & false & false \\ L_{28} & false & W_{28,16} & false & false & false & false \\ L_{29} & false & false & false & false & false & W_{29,18} \\ L_7 & false & false & W_{7,17} & W_{7,5} & W_{7,6} & false \end{array},$$

and:

- $W_{22,15} = W_{17,15} = W_{9,15} = W_{37,15} = W_{32,15} = W_{27,15} =$ "$W_{40,37}$ = " κ - clusters from clustering of type i are received ";
- $W_{23,16} = W_{18,16} = W_{10,16} = W_{38,16} = W_{33,16} = W_{28,16} =$ " n - points between the clusters from clustering of type i are received ";
- $W_{24,18} = W_{19,18} = W_{11,18} = W_{39,18} = W_{34,18} = W_{29,18} =$ " it is necessary other type of clustering ";
- $W_{7,17}$ = " it is necessary new data for clustering ";
- $W_{7,5}$ = " it is necessary new criteria for clustering ";
- $W_{7,6}$ = " it is necessary new method for clustering ".

The tokens in the transition have different marks but we suppose that they are α-tokens (because data, criteria and clustering method entering in the subnet like one token). The tokens that enter in places a_{15}, a_{16}, a_{18} have the following characteristics: *"received clusters* of type *i "* in place a_{15}, *"received n - points between the clusters from clustering of type i "* in place a_{16} and *" need to select other type of clustering"* in place a_{18}.

The tokens that enter in places a_{17}, b_5, c_6 don't obtain new characteristics, respectively: *"needs of new data "* in place a_{17}, *"needs of new criteria "* in place b_5, *" needs of a new method for clustering"* in place c_6.

3 Conclusion

In the presented paper is constructed a generalized net model of the process of selecting a method for clustering. The model describes the different types of clustering. The constructed generalized net is a subnet of the transition Z_6 of the generalized net of the process of clustering. Each one of the transitions Z_1, Z_2, Z_3, Z_4, Z_5, Z_6 in the generalized net can be replaced with subnet again by applying the hierarchical operator H_3. These subnets the authors will construct in the future works.

References

[1] Ahlemeyer-Stubbe A., S. Coleman, *A Practical Guide to Data Mining for Business and Industry*, 2014 John Wiley & Sons, Ltd

[2] Atanassov, K. *Generalized Nets*. World Scientific, Singapore, 1991.

[3] Atanassov, K. *On Generalized Nets Theory*. Prof. M. Drinov Academic Publishing House, Sofia, 2007.

[4] Bramer M., *Principle of Data Mining*, second Edition, Springer London 2007, 2013, ISBN 978-1-4471-4883-8

[5] Bureva, V., Algorithms for associative rule mining, *Management and Education*, Vol. 9, 2013, 121-128 (in Bulgarian).

[6] Bureva, V., E. Sotirova, Generalized net of the process of association rules discovery by Eclat algorithm using weather databases, *14th Int. Workshop on Generalized Nets Burgas*, 29–30 November 2013, 1–10.

[7] Bureva, V., Generalized Net Model of the process of the creating of the associative rules via algorithm Apriori, *Annual of "Informatics" Section Union of Scientists in Bulgaria*, Vol. 5, 2012, 73-83 (in Bulgarian).

[8] Orozova, D., E. Sotirova, Generalized net model of the applying data mining tools, *Proc. of the Tenth International Workshop on Generalized Nets*, Sofia, 2009, 22–26.

[9] Sotirova, E., D. Orozova, Generalized net model of the phases of the data mining process, Developments in Fuzzy Sets, *Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics*. Vol. II: Applications, Warsaw, Poland, 2010, 247–260.

[10] Sotirova. E., K. Dimitrova, R. Papancheva, A Generalized Net Model for Analysis of a Student's Evaluations by Data Mining Techniques in the e-Learning university, *Proc. of the Tenth International Workshop on Generalized Nets*, Sofia, 2009, 41–46.

[11] Zaki. M., Jr. Wagner, *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, May 2014