Generalized Net Model of Cluster Analysis Using CLIQUE: Clustering in Quest

Veselina Bureva[1*], Stanislav Popov[1], Velichka Traneva[2], Stoyan Tranev[3]

[1]*Faculty of Technical Science, Intelligent Systems Laboratory*
University "Prof. Dr. Assen Zlatarov"
Prof. Yakimov Blvd., Burgas 8000, Bulgaria
E-mails: vbureva@btu.bg, stani_popov@yahoo.com

[2]*Faculty of Natural Science*
University "Prof. Dr. Assen Zlatarov"
Prof. Yakimov Blvd., Bourgas 8000, Bulgaria
E-mail: veleka13@gmail.com

[3]*Faculty of Social Science*
University "Prof. Dr. Assen Zlatarov"
Prof. Yakimov Blvd., Bourgas 8000, Bulgaria
E-mail: tranev@abv.bg

[*]*Corresponding author*

Received: August 16, 2018 **Accepted: March 29, 2019**

Published: June 30, 2019

Abstract: *The purpose of the cluster analysis is to find groups of objects with similar characteristics. Different methods are already developed. In the current research work a CLIQUE: Clustering in quest algorithm is investigated. The presented method combines subspace grid-based and density-based techniques to determine clusters of objects. Generalized net of cluster analysis using CLIQUE: Clustering in quest algorithm is constructed. The presented Generalized net can be used for description and monitoring the parallel processes in the cluster analysis.*

Keywords: *Cluster, Cluster analysis, Clique, Generalized nets.*

Introduction

Cluster analysis finds similarities between data according to the characteristics found in the objects and combines similar data into clusters. A cluster is a group of similar objects which are different from the objects included in other clusters. Traditional clustering algorithms often calculate the similarity of objects in the cluster by distance measure. Depending on the approach used to determine the distance between two objects, different cluster analysis algorithms have been developed [1, 5, 12-14]. Depending on the techniques for its realization several variations of cluster analysis exists: partitioning methods, hierarchical methods, grid-based methods, density-based methods, model-based methods and combinations of them. These methods are considered to be fundamental when developing more sophisticated or hybrid models.

In the current research work one of the techniques combining subspace grid-based clustering and density-based cluster analysis is studied. The main steps in the process of detecting groups of objects with similar behavior are: dividing the data space into a finite number of cells, forming a grid-based structure, detecting groups of similar objects, and defining the clusters. Different approaches to grid-based cluster analysis are studied. More common algorithms are STING (Statistical Information Grid) approach, WaveCluster (a Wavelet Based Clustering) approach and CLIQUE: Clustering in Quest [13]. STING (Statistical Information Grid) approach explores statistical information stored in grid cells, WaveCluster clusters objects using a wavelet transform method and CLIQUE: Clustering in quest method represents a combined approach of subspace grid-based and density-based techniques for clustering in a high-dimensional data space. A more thorough study on CLIQUE: Clustering in quest is conducted in the work [1]. CLIQUE: Clustering in quest method automatically identifies the subspaces of the multidimensional data space and allows better clustering than the one in the original space [1, 12, 21].

Generalized net model of the process of clustering in QUEST

CLIQUE: Clustering in quest is introduced by Agrawal et al. [1] and it can be considered as both density-based and grid-based clustering method. CLIQUE: Clustering in quest method automatically identifies subspaces of a high dimensional data space that allows better clustering than the original space. CLIQUE: Clustering in quest method determines clusters by dividing each dimension into $\varnothing$ equal-width intervals and storing those intervals where the density is greater than t as clusters. $\varnothing$ is a previously defined parameter presenting the size of the intervals. Thereafter each dataset of two dimensions is examined: if two intersecting intervals in these two dimensions exist and the density in the intersection of these intervals is greater than t, the intersection is saved as a cluster. The parameter t is previously defined and determines the density of the cluster. The process is repeated for all datasets of all the dimensions. After every step adjacent clusters are replaced by a joint cluster. CLIQUE: Clustering in quest method is insensitive to the order of records in the input and does not presume some canonical data distribution. The selected method scales linearly with the size of input and has good scalability as the number of dimensions in the data increases [10]. The weakness of the algorithm is that the accuracy of the clustering result may be degraded at the expense of simplicity of the method. CLIQUE: Clustering in quest algorithm has the following steps [1, 21]:

- Identify subspaces that contain clusters:
  - Partition the data space and find the number of points that lie inside each cell of the partition.
  - Identify the subspaces that contain clusters using the Apriori principle.
- Identify clusters:
  - Determine the dense units in all subspaces of interests.
  - Determine connected dense units in all subspaces of interests.
- Generate minimal description for the clusters:
  - Determine maximal regions that cover a cluster of connected dense units for each cluster.
  - Determination of minimal cover for each cluster.

The preprocessing step in the algorithm applying is used for determining to the “dirty” data and removing incorrect or null values, performing noise detection, applying data format transformation and etc. The received data will be “cleaned” and valuable for the clustering analysis.

The process of cluster analysis with CLIQUE: Clustering in quest algorithm is modeled using the possibilities of Generalized nets (GNs). The theory of GNs was introduced by Atanassov [2, 3]. GNs are defined in a way that is principally different from the ways of defining the other types of Petri nets. GNs are used for description and modeling of real processes as well as to simulate and control them. They can help us to determine an improvement to the real process.

The constructed here generalized net presents the steps of the CLIQUE: Clustering in quest cluster analysis process. It is a part of a series of models describing data mining processes [4, 6-20].

The constructed GN model of the process of cluster analysis using CLIQUE: Clustering in quest contains 5 transitions and 17 places (Fig. 1). The transitions represent the following processes:

- Z_1 – Multidimensional database;
- Z_2 – Preprocessing step;
- Z_3 – Partition the data space and find the number of points that lie inside each cell of the partition;
- Z_4 – Identify the subspaces that contain groups of objects using the Apriori principle and identify clusters (determine the dense units in all subspaces of interests and connected dense units in all subspaces of interests);
- Z_5 – Generate minimal description for the clusters.

Initially in place L_4 there is one α_1-token. It will be in its own place during the entire time of the GN functioning. It has the following characteristic: "*Multidimensional database*".

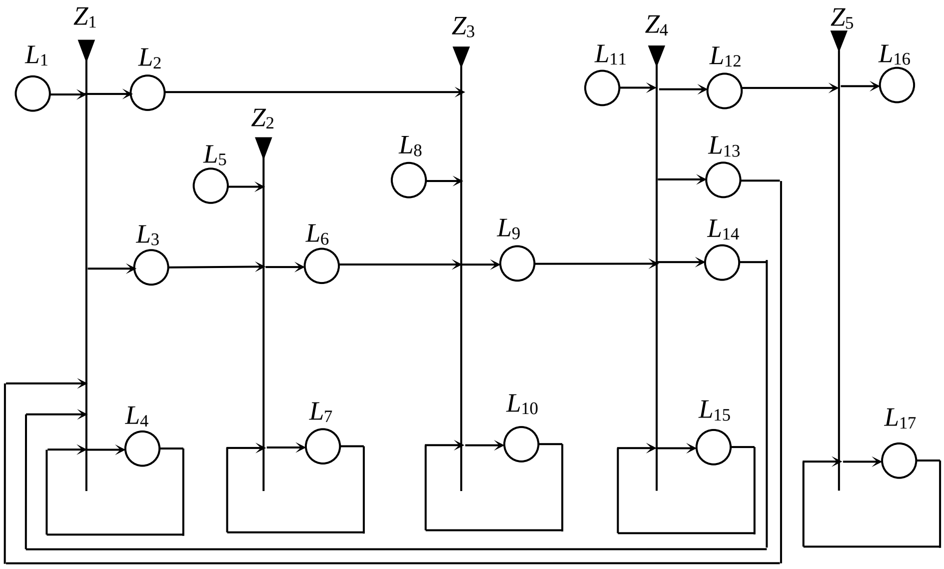


Fig. 1 Generalized net of the cluster analysis process using CLIQUE: Clustering in quest

At the start of the GN's operation there is also one β_1-token that is located in place L_7 with the initial characteristic "*Preprocessing methods*". Preprocessing methods includes techniques for removing the noise of the data; deletion incorrect or null values; data type transformation. The β_1-token in place L_7 generates new β-tokens at certain points in time which can move to place L_6 with the characteristics: "*Preprocessed multidimensional data*".

α_2-token enters the net via place L_1. This token has the initial characteristics: *"Multidimensional data"*.

Transition Z_1 has the following form:

$$Z_1 = \langle\{L_1, L_4, L_{13}, L_{14}\}, \{L_2, L_3, L_4\}, R_1, \vee(L_1, L_4, L_{13}, L_{14})\rangle,$$

where

$R_1 =$

	L_2	L_3	L_4
L_1	*false*	*false*	*true*
L_4	$W_{4,2}$	$W_{4,3}$	$W_{4,4}$
L_{13}	*false*	*false*	*true*
L_{14}	*false*	*false*	*true*

and

- $W_{4,2}$ = "There are selected multidimensional data for performing dividing procedure";
- $W_{4,3}$ = "There are selected multidimensional data for performing preprocessing step";
- $W_{4,4} = \neg (W_{4,2} \wedge W_{4,3})$.

The α-tokens, entering place L_4 do not obtain new characteristics. The α_1-token in place L_4 generates new α-tokens that enter places L_2 and L_3 with the characteristics: "*Selected multidimensional data for performing dividing procedure*" in place L_2 and "*Selected multidimensional data for performing preprocessing step*" in place L_3.

β_2-token enters the net via places L_5. This token has the initial characteristics: "*Preprocessing methods*".

Transition Z_2 has the following form:

$$Z_2 = \langle \{L_3, L_5, L_7\}, \{L_6, L_7\}, R_2, \vee(L_3, L_5, L_7) \rangle,$$

where

$R_2 =$

	L_6	L_7
L_3	*false*	*true*
L_5	*false*	*true*
L_7	$W_{7,6}$	$W_{7,7}$

and

- $W_{7,6}$ = "There are preprocessed multidimensional data";
- $W_{7,7} = \neg W_{7,6}$.

The tokens, entering place L_7 do not obtain new characteristics.

β_3-token enters the net via place L_8. This token has the initial characteristics: "*Parameter for dividing the cells-Ø*".

Transition Z_3 has the following form:

$$Z_3 = \langle \{L_2, L_6, L_8, L_{10}\}, \{L_9, L_{10}\}, R_3, \vee(\wedge(\wedge(L_2, L_8), L_6), L_{10}) \rangle,$$

where

$R_3 =$

	L_9	L_{10}
L_2	*false*	*true*
L_6	*false*	*true*
L_8	*false*	*true*
L_{10}	$W_{10,9}$	$W_{10,10}$

and

- $W_{10,9}$ = “There is a space divided into cells with calculated number of points (that lies inside each cell)”;
- $W_{10,10} = \neg W_{10,9}$.

The tokens, entering place L_{10} do not obtain new characteristics. The token in place L_{10} generates a new token that enters place L_9 with the characteristics: “*Space divided into cells*”.

β_4-token enters the net via place L_{11}. This token has the initial characteristics: “*Density threshold - t*”.

Transition Z_4 has the following form:

$$Z_4 = \langle\{L_9, L_{11}, L_{15}\}, \{L_{12}, L_{13}, L_{14}, L_{15}\}, R_4, \vee(\wedge(L_9, L_{11}), L_{15})\rangle,$$

where

$$R_4 = \begin{array}{c|cccc} & L_{12} & L_{13} & L_{14} & L_{15} \\ \hline L_9 & false & false & false & true \\ L_{11} & false & false & false & true \\ L_{14} & false & false & false & true \\ L_{15} & W_{15,12} & W_{15,13} & W_{15,14} & W_{15,15} \end{array}$$

and

- $W_{15,12}$ = “There are identified clusters (determine the dense units in all subspaces of interests and connected dense units in all subspaces of interests) after applying the Apriori principle”;
- $W_{15,13}$ = “There are cells which need further processing after applying Apriori property”;
- $W_{15,14}$ = “There are cells that are not satisfied the density threshold”;
- $W_{15,15} = \neg (W_{15,12} \wedge W_{15,13} \wedge W_{15,14})$.

The tokens, entering place L_{17} do not obtain new characteristics. The token in place L_{17} generates new tokens that enter places L_{12}, L_{13} and L_{14} with the characteristics: “*Identified clusters (determine the dense units in all subspaces of interests and connected dense units in all subspaces of interests) after applying the Apriori principle*” in place L_{12}, “*Cells which need further processing after applying Apriori property*” in place L_{13} and “*Cells that are not satisfied the density threshold*” in place L_{14}.

Transition Z_5 has the following form:

$$Z_5 = \langle\{L_{12}, L_{17}\}, \{L_{16}, L_{17}\}, R_5, \vee(L_{12}, L_{17})\rangle,$$

where

$$R_5 = \begin{array}{c|cc} & L_{16} & L_{17} \\ \hline L_{12} & false & true \\ L_{17} & W_{17,16} & W_{17,17} \end{array}$$

and

- $W_{17,16}$ = “There is a minimal description for the clusters”;
- $W_{17,17} = \neg W_{17,16}$.

The tokens, entering place L_{17} do not obtain new characteristics. The α-token in place L_{17} generates a new token that enters place L_{16} with the characteristic: "*Generated minimal description for the clusters*".

Conclusion

In the current paper, a generalized net of the cluster analysis process by the CLIQUE: Clustering in quest algorithm is constructed. CLIQUE: Clustering in quest is a subspace grid-based and density-based algorithm for performing cluster analysis in multidimensional space. The presented Generalized net can be used for description and monitoring the parallel processes in the cluster analysis.

Acknowledgements

The authors are thankful for the support provided by the Bulgarian National Science Fund under Grant Ref. No. DFNI-DN-02-10/2017 "New Instruments for Knowledge Discovery from Data and their Modelling" and the project of University "Prof. Dr. Assen Zlatarov" under Ref. No. NIH-401/2017 "Modern Methods of Optimization and Business Management".

References

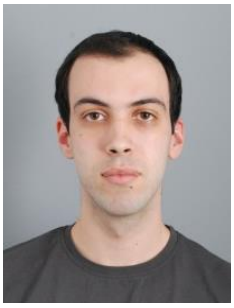
1. Agrawal R., J. Gehrke, D. Gunopulos, P. Raghavan (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, Proceeding of the SIGMOD International Conference on Management of Data, 27(2), 94-105.
2. Atanassov K. (1991). Generalized Nets, World Scientific, Singapore.
3. Atanassov K. (2007). On Generalized Nets Theory, Prof. M. Drinov Academic Publishing House, Sofia.
4. Atanassov K. (2016). Generalized Nets as a Tool for the Modelling of Data Mining Processes, Innovative Issues in Intelligent Systems, Studies in Computational Intelligence, 623, 161-215.
5. Bhoi A. K. (2017). Classification and Clustering of Parkinson's and Healthy Control Gait Dynamics Using LDA and K-means, International Journal Bioautomation, 21(1), 19-30.
6. Bureva V., E. Sotirova, K. Atanassov (2014). Hierarchical Generalized Net Model of the Process of Clustering, Issues in Intuitionistic Fuzzy Sets and Generalized Nets, Vol. 1, Warsaw School of Information Technology, 73-80.
7. Bureva V., E. Sotirova, K. Atanassov (2014). Hierarchical Generalized Net Model of the Process of Selecting a Method for Clustering, 15th International Workshop on Generalized Nets, 16 October 2014, Burgas, 39-48.
8. Bureva V., E. Sotirova, S. Popov, D. Mavrov, V. Traneva (2017). Generalized Net of Cluster Analysis Process Using STING: A Statistical Information Grid Approach to Spatial Data Mining, Lecture Notes in Computer Science, 10333, 226-238.
9. Bureva V., S. Popov, E. Sotirova, B. Miteva (2017). Generalized Net of the Process of Hierarchical Cluster Analysis, Assen Zlatarov University Annual, Technical and Natural Sciences, XLVI(1), 107-111.
10. Erbakanov L., K. Atanassov, S. Sotirov (2015). Generalized Net Model of a Body Temperature Data Logger Embedded System, International Journal Bioautomation, 19(2), 237-244.
11. Georgieva V. (2017). Generalized Net Model of Mechanical Wastewater Pre-treatment, International Journal Bioautomation, 21(1), 133-144.
12. Guha S., R. Rastogi, K. Shim (1998). CURE: An Efficient Clustering Algorithm for Large Databases, SIGMOD'98, 73-84.
13. Jain K. A., R. C. Dubes (1988). Algorithms for Clustering Data, Printice Hall.

14. Kaufman L., P. J. Rousseeuw (1990). Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons.
15. Ribagin S., O. Roeva, T. Pencheva (2016). Generalized Net Model of Asymptomatic Osteoporosis Diagnosing, 2016 IEEE 8th International Conference on Intelligent Systems, IS 2016, 604-608.
16. Roeva O., A. Shannon, T. Pencheva (2012). Description of Simple Genetic Algorithm Modifications Using Generalized Nets, Proceedings of the 6th IEEE International Conference Intelligent Systems, 178-183.
17. Simeonov S., V. Atanassova, E. Sotirova, N. Simeonova, T. Kostadinov (2018). Generalized Net of a Centralized Embedded System, Uncertainty and Imprecision in Decision Making and Decision Support: Cross-Fertilization, New Models and Applications, IWIFSGN 2016, Advances in Intelligent Systems and Computing, 559, 299-304.
18. Sotirov S., E. Sotirova, M. Werner, S. Simeonov, W. Hardt, N. Simeonova (2016). Ituitionistic Fuzzy Estimation of the Generalized Nets Model of Spatial-temporal Group Scheduling Problems, Imprecision and Uncertainty in Information Representation and Processing, Studies in Fuzziness and Soft Computing, 332, 401-414.
19. Sotirov S., M. Werner, S. Simeonov, W. Hardt, E. Sotirova, N. Simeonova (2014). Using Generalized Nets to Model Spatial-temporal Group Scheduling Problems, Issues in Intuitionistic Fuzzy Sets and Generalized Nets, Vol. 11, 42-54.
20. Sotirova E., D. Orozova (2010). Generalized Net Model of the Phases of the Data Mining Process, Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics, II: Applications, Warsaw, Poland, 247-260.
21. Yadav J., D. Kumar (2014). Subspace Clustering Using CLIQUE: An Exploratory Study, International Journal of Advanced Research in Computer Engineering & Technology, 3(2), 372-378.

Assist. Prof. Veselina Bureva, Ph.D.
E-mail: vbureva@btu.bg

Veselina Bureva graduated in Communication and Information Systems at "Konstantin Preslavsky" University of Shumen. She completed Ph.D. in Computer Systems and Technologies in "Prof. Dr. Assen Zlatarov" University, Burgas. Since 2011, she has been working in the Computer Systems and Technologies Department of the Faculty of Technical Sciences at the "Prof. Dr. Assen Zlatarov" University. Her current research interests are in the field of data mining, generalized nets modeling, and etc.

Stanislav Popov, Ph.D. Student
E-mail: stani_popov@yahoo.com



Stanislav Popov holds his Bachelor's Degree in International Economic Relations at University of Economics, Varna, Bulgaria. He has two Master's degrees of International Business at University of Economics, Varna and Computer System and Technologies at "Prof. Dr. Assen Zlatarov" University, Burgas. At the moment he is a Ph.D. student in Computer Systems and Technologies at "Prof. Dr. Assen Zlatarov" University. His interests are related to data mining, generalized nets modeling, international business process, etc.

Assist. Prof. Velichka Traneva, Ph.D.
E-mail: veleka13@gmail.com

Assistant Professor V. N. Traneva has Ph.D. in Informatics of Bulgarian Academy of Sciences, holds a Master's degree in Finance and Mathematics and professional qualification Manager at Sofia University "Saint Kliment Ohridski", Bulgaria. Shee is a co-author of numerous publications in Mathematics and Informatics. Her interests are in the field of mathematical modelling, optimization and informatics.

Assist. Prof. Stoyan Tranev, Ph.D.
E-mail: tranev@abv.bg

Assistant Professor S. T. Tranev has Ph.D. in Organization and Management of Production of University "Prof. Dr. Assen Zlatarov", Bulgaria, holds a Master's degree in Marketing and Management and professional qualification Mediator in Conflictology to International Academy under Informatization, Institute under Conflictology. He is a co-author of numerous publications in Economics. His interests are in the field of management and conflictology.

© 2019 by the authors. Licensee Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).