Olga PILIPCZUK[1], Galina CARIOWA[2]

[1]Faculty of Economics and Management, University of Szczecin, [2]ZFaculty of Computer Science and Information Technology, West Pomeranian University of Technology

doi:10.15199/48.2015.11.47

The new module for rules discovering and visualization for NovoSpark® Visualizer software

Abstract. *In this paper we present the new rough sets module for NovoSpark® Visualizer (NV) software. We describe the NV system architecture and the place of the new module in it. We also present the procedure of rough sets analysis with NV software. In addition an example of rules discovering and visualization is provided to evaluate the proposed module. The results show that useful rules are discovered efficiently from the data set.*

Streszczenie. *W artykule zaprezentowano projekt nowego modułu do automatyzacji teorii zbiorów przybliżonych dla oprogramowania NovoSpark® Visualizer (NV). Opisano architekturę systemu oraz wskazano miejsce nowego modułu. Ponadto zaprezentowano przebieg procedury analizy i wizualizacji zbiorów przybliżonych w systemie. Przedstawiono przykład odkrywania i wizualizacji reguł za pomocą opracowanej procedury. W wyniku przeprowadzenia eksperymentu udało się otrzymać szereg użytecznych reguł decyzyjnych.* ***(Nowy moduł odkrywania i wizualizacji reguł dla systemu NovoSpark® Visualizer).***

Keywords: rough sets, visualization, rule
Słowa kluczowe: zbiory przybliżone, wizualizacja, reguły

Introduction

Artificial intelligence methods produce enormous amounts of rules. The problem of rules discovering from large databases is one of the most challenging research subjects and the topic of numerous studies. However, rules discovering generally results in a large number of found rules, and analysts must identify manually the useful ones. When the number of conditions in a decision rule increases, and the overall number of rules is approximately 20-50, it is extremely difficult to analyze and extract the knowledge [1]. “Sifting manually through large sets of rules is time consuming and strenuous” [2].

In this situation, methods and tools that enable rules visualization are of great importance. Visualization can help to improve the intelligibility of the large rule sets, carry out the manipulation inside them, and discover the rules. “However, most rule visualization techniques are still falling short when it comes to a large number of rules” [2].

In this paper we present the project of the NV rough sets module for rules analysis and visualization.

Related Work

In recent years, several visualization techniques have been developed, and various new techniques to visualize the classification rules or association rules and decision rules have been created.

One of the most popular techniques for rules visualization is the scatter plot [3] and its new version - two-key plots [4].

Graph-based interactive visualization and exploration platforms for networks and graphs have been presented in different works [5-10].

Another way of representing the rule sets is the matrix-based approach as a directed 2D graph and its 3D interpretation. In their paper [2], M. Hahsler and S. Chelluboina presented “most interesting rules” by coloring and positioning them in a special way in a grouped matrix. In [11], the authors presented a version of the 2D matrix-based visualization technique in which the interest measure is represented by color shading of squares at the intersection. An alternative method is to use 3D bars at the intersections [12].

Another popular method - parallel coordinates - could be used to visualize different types of rules [13, 14]. A fascinating technique features Hofmann and Wilhelm's mosaic plots, which are inspired from parallel coordinates [15]. The authors also introduced double-decker plots to visualize a single association rule [16].

A creative proposition was made in work [17]. The authors present the rules as the spheres put on top of cones in the landscape, thus obtaining three straightforward graphical characteristics to represent quality measures: sphere diameter, cone height, and color.

The relatively new concept of rules visualization is to combine the different visualization techniques. The new visualization technique proposed by Włodyka et al. is based on two methods of the rules visualization: decision trees (AQDT-2 algorithm) and rule-diagrams. Grzegorz Ilczuk and Alicja Wakulicz-Deja also proposed a rules visualization method based on iDecision trees [18]. The SARV tool provides the three synchronized views: the matrix view featuring a global view of the rules, the graph view illustrating relationships of the rules, and the detailed view of selected rules or items [19].

A more advanced technique is the Virtual Reality approach, which is introduced for the problem of decision rules generated by inductive methods, rough set algorithms, and others [20]. More propositions for rules visualization are presented in other works [21-28].

However most of the visualization techniques have different weaknesses. The most widespread problem involves the processing of large data sets. Several visualizations are overloaded with labels and crossing lines. Some of them are useful just for singe rule analysis.

NovoSpark Visualizer Software

The NovoSpark® Visualizer is a tool for the analysis and visualization of multidimensional data. Currently, Visualizer supports traditional methods for multidimensional data analysis for classification of parameters, observations and individual data sets: discriminant analysis, cluster analysis, regression analysis, factor analysis, self-organizing maps, and descriptive statistics.

NovoSpark® Visualizer is a desktop application implementing the multi-SDI windowing paradigm (see Fig. 1) [29]. Each window holds a separate project that can be persisted to a project file and manipulated independently from other project windows. At the core of each

NovoSpark® Visualizer project is an OpenGL-based rendering component that shows a composite multidimensional image of all datasets loaded into the project window. The renderer comes with a number of data visualization and transformation options. The structure and contents of each dataset can be viewed and/or modified in a window hosting a powerful FlexCell grid optimized to work with potentially large volumes of data.

NovoSpark® Visualizer hosts a framework of pluggable data analysis and visualization modules that can be downloaded and installed separately from the main application [29].

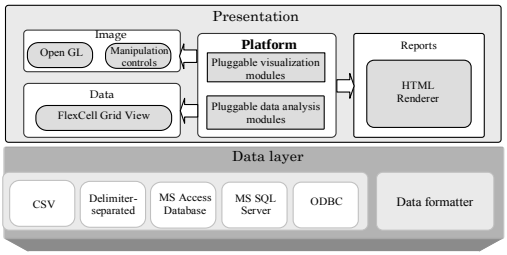


Fig.1. The NV architecture (Source: [29])

This feature makes it easy to plug the new module into the system platform. The underlying data layer forms a basis for retrieval and formatting of datasets loaded from external data sources [29].

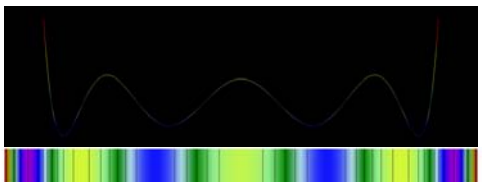


Fig.2. The observation of A-point with its spectrum

The NovoSpark® Visualizer method is based on the following set of formulas [30]:

1) For the a point-observation *A* in *N*-dimensional affine point-vector space R_N of the originals $A=(a_0,a_1...a_{N-1})$ and form linear combination $f_A(t)$ of functions $\{P_i(t)\}_0^\infty$ by using the following rule [30]:

(1) $$f_A(t)=\sum_{i=0}^{N-1} a_i P_i(t),$$

where: $P_i(t)$ - orthogonal polynomials with weight 1 defined on the segment *t* = [0, 1].

This point can be "painted" in accordance with the function values. It is called by authors a “spectrum” of the multidimensional point-observation (see Fig. 2) [30].

2) For multidimensional interval of the point *X* at a position z [30, 31]:

(2) $$X=X(z)\leftrightarrow f_X(t)=\sum_{i=0}^{N-1} x_i(z)P_i(t)=f_X(t,z),$$

where: $x_i(z)=a_i+z(b_i-a_i)$.

Radius vector: $p_B=vector(b_0,b_1...b_{N-1})$ for any point *X* belonging to the segment *AB*: $p_X=vector(x_0,x_1...x_{N-1})$ satisfies the following equation [30, 31]:

(3) $p_X=p_A+z(p_B-p_A)=p_X=vector(x_0(z),x_1(z)...x_{N-1}(z))$,

where: $z\in[0,1]$.

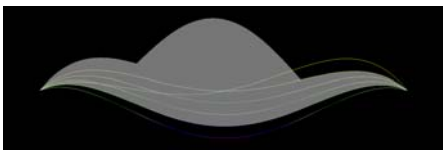


Fig.3. The “cloud” image example

The image of a multidimensional interval is a two-dimensional region between the “minimum” and “maximum” images [30] in the coordinate system {f, t} (see Fig.3). The authors term this a "cloud" of a multidimensional interval. Boundaries of the cloud are obtained from a linear combination of separate images from the coordinate space of the originals [30].

The rough sets module for NV tool

The proposed module is pluggable, as are the other data analysis modules, and it is integrated with the existing visualization modules. The system will generate a report with the main rough sets measures.

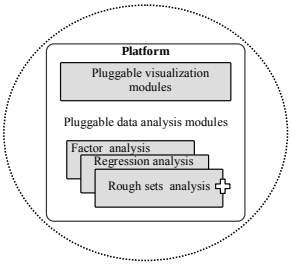


Fig.4. Rough sets module in NV platform architecture

The rules are treated as function-curves in a multidimensional space. The procedure of rules discovering is performed as follows:

1. Introduction or downloading of data into the system.
2. Creating the global view using parallel coordinates.
3. Finding reducts by excluding columns from the dataset and analyzing the image changes.
4. Grouping elementary sets - creating the image of a "cloud" by adjusting the confidence interval.
5. Elementary sets comparison using spectrum images.
6. Rules creation and attractiveness analysis [32].

Experiment

During the experiment, the ability to create useful rules based on examples of numerical data was tested.

Table 1. Information table

N	C_1	C_2	C_3	C_4
1	4	2	1	1
2	3	1	2	1
3	1	1	1	1
4	5	3	3	1
5	4	3	3	1
6	4	2	3	1
7	4	3	2	1
8	3	2	1	1
9	1	3	3	1
10	4	2	3	1
11	2	2	2	1
12	2	1	2	1
13	3	2	1	1
14	4	3	2	1
15	3	2	2	1
16	5	1	1	1
17	5	2	2	1
18	5	3	3	1
19	2	2	3	1
20	1	3	3	1

21	4	2	2	1
22	4	1	1	1
23	5	2	3	1
24	3	1	2	1
25	2	1	2	1
26	3	2	2	1
27	4	1	3	1
28	5	1	2	1
29	1	3	3	1
30	1	2	3	1
31	4	1	2	1
32	3	2	2	1
33	2	2	3	1
34	3	1	1	1
35	2	3	3	1
36	5	1	3	1
37	5	3	1	1
38	3	3	3	1
39	4	2	2	1
40	4	1	3	1
41	4	2	1	1
42	3	1	1	1
43	1	1	1	1
44	1	3	3	1
45	1	3	2	1

The data from Tab. 1 was introduced into the system, where C_1-C_4 – are condition attributes. First, we created a global view of the decision situation (see Fig. 5). The first parameter has the highest value is shown in red. Analyzing the portion of the image representing the second and third parameter, we note that the overall level of ratings is similar. The last, fourth parameter, is a constant; it's value shown in purple.

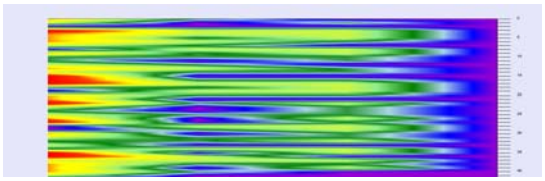


Fig.5. The data set global view

The next step is the reducts identification done by excluding columns from the entire dataset. The attributes elimination causes changes in the shape of the "cloud" image or in the number of curves. The more the shape changes, the greater the importance of the attribute.

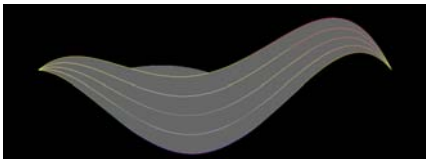
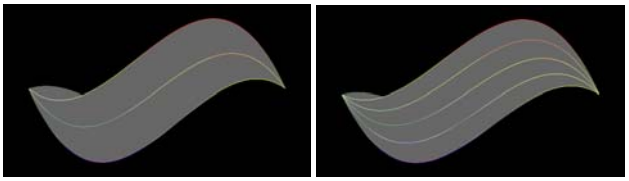
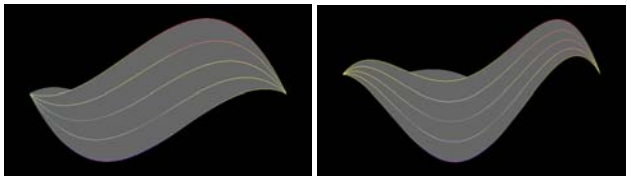


Fig.6 The "cloud" image for entire dataset

Reducing attribute C_1, we notice that the image changes significantly, which means that the attribute is not redundant (see Fig. 7a).



a) b)



c) d)

Fig.7. a) C_1 reduction b) C_2 reduction c) C_3 reduction d) C_4 reduction

Reducing attribute C_2, we notice that the image is changing, but to a lesser extent than the previous one (Fig. 7b). This means that the attribute is not redundant, but its importance is less than the importance of attribute C_1.

Reducing attribute C_3, we notice that the picture is changing; it adopts a similar form as the previous one (see Fig. 7c). This means that the attribute is not superfluous, but its importance is less than the importance of C_1. An attempt to reduce the attribute C_4 ended successfully, as clearly shown in Fig. 7d.

The next step is the rules grouping, using the cloud images. During the experiment, the 1.0 confidence interval was applied. As a result, the following collections were received: E1 {4,16,17,18, 23, 28, 36, 37}; E2 {21, 39}; E3 {15, 26, 32}; E4 {5, 7, 14, 35, 38}; E5 {1, 8, 13, 41}; E6 {11}; E7 {2, 12, 22, 24, 25, 27, 31, 34, 40, 42}; E8 {3, 9, 20, 29, 30, 43, 44, 45}.

Each of these sets can further be analyzed to determine the similarity between rows. Figure 8 shows the "spectrum" image of E8 {3, 9, 20, 29, 30, 43, 44, 45}.

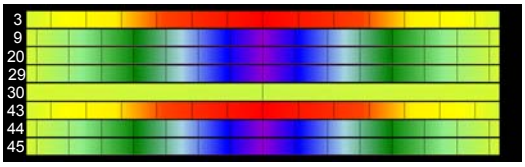


Fig.8. E8 set "spectrum" image

Rows 3 and 43 are identical, as well as the rows from group 9, 20, 29, 44, and 45. An interesting case is row number 45, which is identical to this group. Although the value of C_3 is equal to 2, not 3 as in the other rows within this group (9, 20, 29, 44), due to the large value of the support measure system immediately shows that we can create a rule, without the risk of error occurring. Similarly, we can analyze other groups of rows.

As a result, we obtain the following rules: R1 {3, 43}, R2 {9, 20, 29, 44, 45}, R3 {12, 25}, R4 {30}, R5 {11}, R6 {19, 33}, R7 {35}, R8 {34, 42}, R9 {2, 24}, R10 {8, 13}, R11 {15, 26, 32}, R12 {38}, R13 {22}, R14 {31}, R15 {27, 40}, R16 {1,41}, R17 {4, 18}.

The next step is the addition of decision attribute (d). We received the following results:

Table 2. Decision rules with their confidence

N	C_1	C_2	C_3	C_4	d	confidence
R1	4	2	1	1	2	1
R2	1	1	1	1	2	1
R3	5	3	3	1	1	1
R4	2	1	2	1	2	1
R5	3	2	2	1	2	0,77
R6	1	3	3	1	2	0,8
R7	4	1	3	1	1	1
R8	3	1	1	1	2	1
R9	1	3	2 or 3	1	2	0,8

Results and Discussion

After conducting a series of experiments, we found that the choice of the interval depends on the size of the

universe. With more lines, the confidence interval should be closer to 1.0. For example, when we chose an interval of 3.0, created groups were too large and we had problems with their analysis. These groups are: {4,16,17,18, 23, 28, 36, 37}; {5,6, 7, 10, 14, 21, 22, 27, 31, 39, 40, 41}; {2, 8, 13, 15, 24, 26, 32, 34, 38, 42}; {11, 12, 19, 25, 33, 35}; {3, 9, 20, 29, 30, 43, 44, 45}.

Therefore, the confidence interval was reduced to 1.0. In this case we execute the rules search mechanism from general to specific, decreasing and gradually narrowing the search range and dividing the row groups into subgroups.

The nine useful rules were obtained. Here we treated the rule with the confidence ratio> 0.5 as useful.

We concluded that the system allows for visualization of the large number of rows without loss of transparency.

Conclusion

Experiments suggest that the visualization algorithms used in NovoSpark software can be successfully applied during the rules discovering using rough sets theory. The advantages of the designed module are ease of implementation and the possibility of rough set theory combination with other analyses. “Global view” images allow a comprehensive look at the dataset; “cloud” images help to group rows; “spectrum” image show similarity, which creates the possibility of quick rules discovery. Created rules can be stored in the system in the form of images.

The disadvantage of the above visualization is the lack of possibility to include information about the rules attractiveness measures directly on the image. However, a separate image can be created to analyze the measures. We also need to create a new table (Table 2) to analyze the main rules attractiveness measures such as coverage, support, and usability.

REFERENCES

[1] Wlodyka A., Mlynarski R., Ilczuk G., Pilat E., Kargul W. Visualization of Decision Rules – from the Cardiologist's Point of View, *Proc. Conference Computers in Cardiology*, (2008), 645-648

[2] Hahsler M., Chelluboina S., ArulesViz - Visualizing Association Rules, R package version 0.1-1. 10, (2011)

[3] Bruzzese D., Davino C., Visual Mining of Association Rules, in *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, Springer-Verlag, (2008), 103-122

[4] Unwin A., Hofmann H., Bernt K., The Two Key Plot for Multiple Association Rules Control, *Proc. of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, Springer-Verlag, (2001), 472-483

[5] Klemettinen M., Mannila H., Ronkainen P., Toivonen H., Verkamo A. I., Finding Interesting Rules from Large Sets of Discovered Association Rules, *CIKM*, (1994), 401-407

[6] Rainsford C. P., Roddick J. F., Visualisation of Temporal Interval Association Rules, *Proc. of the Second International Conference on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineering, and Intelligent Agents*, 2000, 91-96

[7] Buono P., Costabile M. F., Visualizing Association Rules in a Framework for Visual Data Mining, *From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments*, (2005), 221-231

[8] Ertek G., Demiriz A., (2006), A Framework for Visualizing Association Mining Results, *ISCIS*, 593-602

[9] Blanchard J., Guillet F., Briand H. Interactive visual exploration of association rules with rule-focusing methodology, *Knowledge and Information Systems*, 13 (2007), No. 1, 43-75

[10] Bastian M., Heymann S., Jacomy M., Gephi: An Open Source Software for Exploring and Manipulating Networks, *ICWSM 8* (2009), 361-362

[11] Ong K.-H., leong Ong K., Ng W.-K., Lim E.-P., CrystalClear: Active Visualization of Association Rules, *ICDM'02 International Workshop on Active Mining*, (2002)

[12] Wong P.C, Whitney P., Thomas J. Visualizing association rules for text mining, *Proc. of the 1999 IEEE symposium on information visualization*, IEEE Computer Society, (1999), 120–123

[13] Han J., An A., Cercone, N., CViz: An Interactive Visualization System for Rule Induction, *LNCS*, (2000), 214-226

[14] Yang L., Pruning and Visualizing Generalized Association Rules in Parallel Coordinates, *Knowledge and Data Engineering*, 17 (2005), 60-70

[15] Hofmann H, Wilhelm A., Visual comparison of association rules, *Comp Stat*, 16 (2001), No.3, 399–415

[16] Hofmann H., Siebes A., Wilhelm A. F. X., Visualizing Association Rules with Interactive Mosaic Plots, in *KDD*, (2000), 227-235

[17] Blanchard J, Fabrice Guillet, Henri Briand Exploratory Visualization for Association Rule Rummaging, *Proc. of the 4th International Workshop on Multimedia Data Mining MDM/KDD* (2003), 107-114

[18] Ilczuk G., Wakulicz-Deja A., Visualization of Rough Set Decision Rules for Medical Diagnosis Systems, *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, *LNCS*, Volume 4482 (2007), 371-378

[19] Sekhavat Y. A., Hoeber O., Visualizing Association Rules Using Linked Matrix, Graph, and Detail Views, *Int. J. of Intelligence Science*, (2013), No.3, 34-49

[20] Valdes J., Virtual Reality Representation of Information Systems and Decision Rules: An exploratory technique for understanding data and knowledge structure. Rough Sets, *Fuzzy Sets, Data Mining, and Granular Computing LNCS*, 2639, (2003), 615-618

[21] Jiang B., Han Ch., Hu X., A finite ranked poset and its application in visualization of association rulet, *GrC 2008*, 2008.

[22] Berrado A., Runger G. C., Using metarules to organize and group discovered association rules, *Data Min. Knowl. Disc.*, 14 (2007), 409-431

[23] Berardi M., Appice A., Loglisci C., Leo P. Supporting Visual Exploration of Discovered Association Rules Through Multi-Dimensional Scaling. ISMIS 2006, *LNAI* 4203 (2006), 369-378.

[24] Carson Kai-Sang Leung, Pourang P. Irani, and Christopher L. Carmichael. FIsViz: A Frequent Itemset Visualizer. PAKDD 2008, *LNAI* 5012 (2008), 644-652

[25] Techapichetvanich K., Datta A., VisAR: A New Technique for Visualizing Mined Association Rules, ADMA 2005, *LNAI* 3584 (2005), 88-95

[26] Marghoubi R., Boulmakoul A., Zeitouni K., The Use of the Galois lattice for the extraction and the visualization of the spatial association rules, *Signal Processing and Information Technology*, (2006), 606-611

[27] Yahia S., Nguifo E., Contextual generic association rules visualization using hierarchical fuzzy meta-rules, *Proc. of Fuzzy Systems*, (2004), No. 1, 227-232

[28] Herawan T., Yanto I.T.R., Deris M. M., SMARViz: Soft Maximal Association Rules Visualization, IVIC 2009, *LNCS* 5857 (2009), 664-674

[29] Pilipczuk O., Shamroni D., Podstawowe aspekty tworzenia systemów grafiki kognitywnej, *Problemy Zarządzania*, 07 (2012), No.10 (3), 248-261

[30] Eidenzon D., Shamroni D., Volovodenko V., Method and System for Multidimensional Data Visualization, LAP Lambert Academic Publishing, (2013)

[31] Eidenzon D., Pilipczuk O., Multidimensional data visualization, *Encyclopedia of Information Science and Technology*, IGI-Global, Hershey, (2014), 1600-1610

[32] Pawlak Z., Skowron A., Rudiments of Rough Sets, *Inform. Sciences*, 177, (2007), No.1, 3-27

***Authors**: dr Olga Pilipczuk, University of Szczecin, Institute of IT in Management, 64 Mickiewicza Str., 71-101, Szczecin, E-mail: olga.pilipczuk@wneiz.pl; dr inż. Galina Cariowa, West Pomieranian Uniersity of Technology,, Department of Multimedia Systems, 49 Żołnierska Str., 70-210, Szczecin, E-mail: gcariowa@wi.zut.edu.pl.*