Tenth Int. Workshop on Generalized Nets
Sofia, 5 December 2009, 22-26

Generalized Net Model of the Applying Data Mining Tools

Daniela Orozova[1] and Evdokia Sotirova[2]
"Prof. Asen Zlatarov" University, Bourgas 8010, Bulgaria
e-mails: *orozova@bfu.bg*, *esotirova@btu.bg*

Abstract: In the paper is constructed a Generalized Net (GN) model of the applying data mining tools, which provides a compact representation of the discovered patterns and allows the model application to new amounts of data. The opportunity of using GNs as a tool for modeling such processes is analyzed as well.

Keywords: Data mining, Generalized nets.

1 Introduction

Data mining is the process of discovery of hidden patterns and dependencies in data (see [3, 4, 5, 6, 7, 8, 10]).

Data mining is an interactive process that starts with understanding and definition of a problem for solving, and finishes with the analysis of the obtained results and the strategy for their practical utilization.

Data mining commonly involves four classes of tools [3]:

- Classification - Arranges the data into predefined groups. For example an email program might attempt to classify an email as legitimate or spam. Common algorithms include Decision Tree Learning, Nearest neighbor, naive Bayesian classification and Neural network.
- Clustering - Is like classification but the groups are not predefined, so the algorithm will try to group similar items together.
- Regression - Attempts to find a function which models the data with the least error.
- Association rule learning - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as "market basket analysis".

In [9] a GN model is constructed of the Cross-Industry Standard Process Model of Data Mining, which provides a compact representation of the discovered patterns and allows the model application to new amounts of data (Figure 1). It consist of seven transitions which represent respectively:

- Z_1 and Z_2 – Problem Understanding Phase;
- Z_3 and Z_4 – Data Understanding Phase;
- Z_5 – Data Preparation Phase;

- Z_6 – Modeling Phase;
- Z_7 – Evaluation Phase;
- Z_8 – Deployment Phase.

In this paper the GN model of the applying data mining tools is constructed. It is a subnet of the GN net form [9] and corresponds to the transition Z_6. The transition Z_6 from [9] can be changed by a hierarchical operator H_3 (see [1, 2]) with a present sub-GN model.

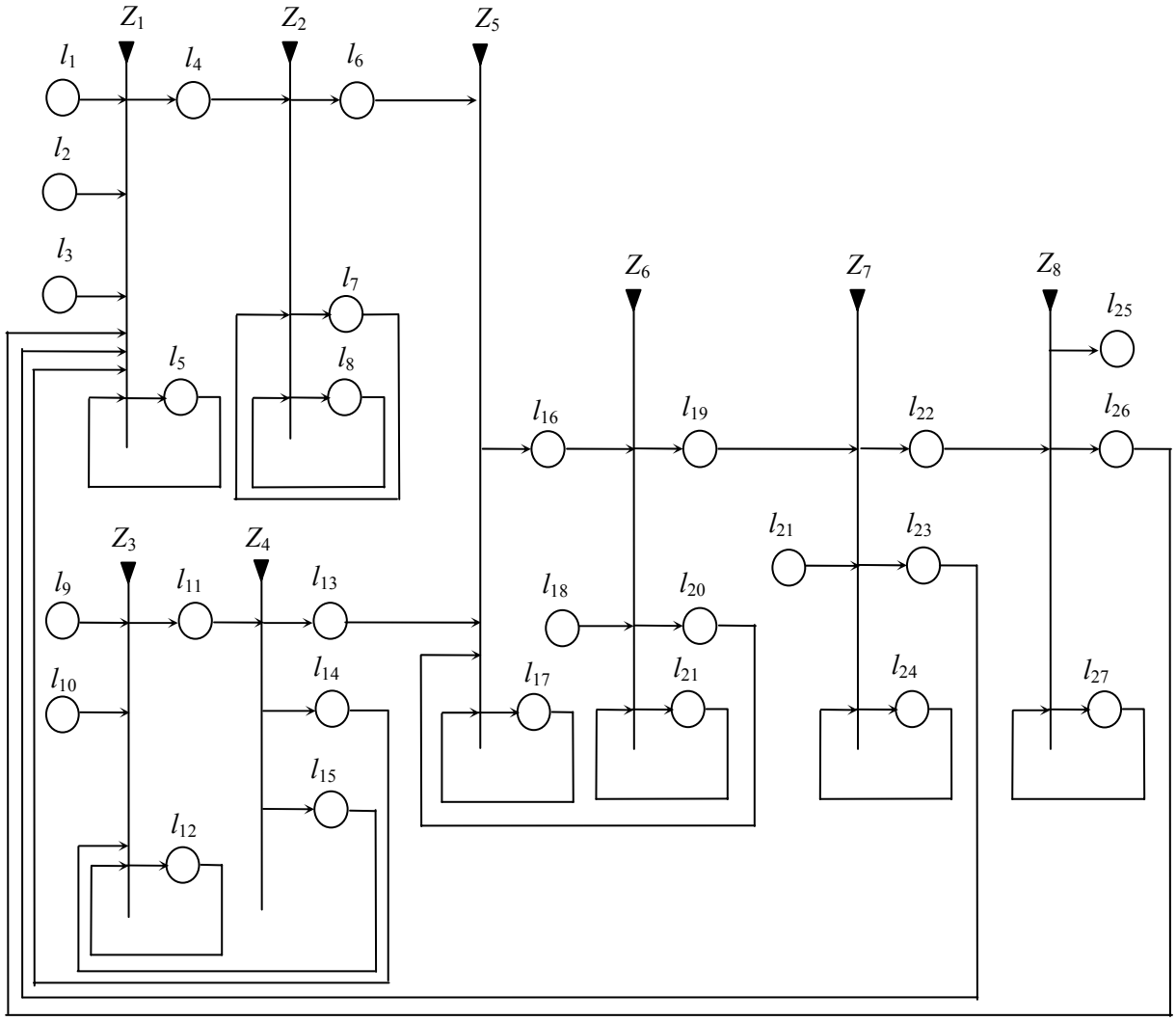

Figure 1: Generalized net model of the CRISP-DM

2 Generalized net model

The GN-model of the applying data mining tools (see Fig. 2) contains 5 transitions and 21 places, collected in tree groups and related to the tree types of the tokens that will enter respective types of places:

- α-tokens and l-places represent the process of the applying the data mining tool,
- β-tokens and t-places represent the criteria for the restricting data mining tool and choosing the property data mining tool.

For brevity, we shall use the notation α- and β-tokens instead of α_i- and β_j-tokens, where *i, j* are numerations of the respective tokens.

Initially, there is one β_0-token that is located in place t_6 with initial characteristic *"Current data mining tools"*. In the next time-moments this token is split into two. One of them, let it be the original β-token, will continue to stay in place t_6, while the other β-tokens will move to transition Z_5 passing via transition Z_2.

The α_0 and α_1-tokens with characteristics *"Initial hypothesis"* and *"Initial data"* enter the net via places l_0 and l_1 respectively.

Tokens β_1 and β_2 enter the net via places t_0 and t_1 respectively. These tokens have initial characteristics *"New data mining tool"* in place t_0, and *"Criteria for the restricting data mining tool"* in place t_1.

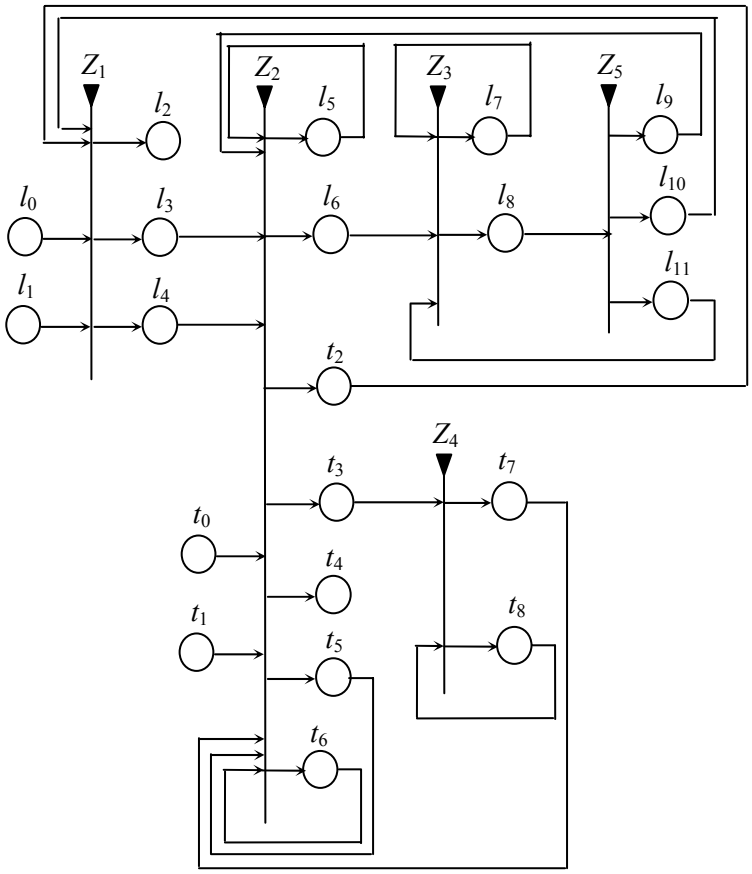

Figure 2: Generalized net model of the applying data mining tools

The forms of the transitions are the following.

$$Z_1 = <\{l_0, l_1, l_{10}, t_2\}, \{l_2, l_3, l_4\},$$

	l_2	l_3	l_4
l_0	*false*	*false*	*true*
l_1	*false*	*false*	*true*
l_{10}	$W_{10,2}$	$W_{10,3}$	*false*
t_2	*false*	*true*	*false*

$$, \vee(\wedge(l_0, l_1), l_{10}, t_2)>,$$

where:

- $W_{10,2}$ = "Data mining tools are applied and evaluated",
- $W_{10,3} = \neg W_{10,2}$.

On the first activation of the transition Z_1 the α_0 and α_1-tokens that enter place l_4 (from places l_0 and l_1) merge in a new α-token with characteristic *"Initial hypothesis, initial data"*.

On the next activation of the transition Z_1 the α-token that enters place l_3 (from places t_2) obtains characteristic *"Goal, data mining tools"*.

On the next activation of the transition Z_1 the α-tokens can enter places l_2 or l_3.The characteristic of the token that enters place l_3 is mentioned above, and the α-token that enters place l_2 do not obtains new characteristic.

$$Z_2 = <\{l_3, l_4, l_5, l_9, t_0, t_1, t_5, t_6, t_7\}, \{l_5, l_6, t_2, t_3, t_4, t_5, t_6\},$$

	l_5	l_6	t_2	t_3	t_4	t_5	t_6
l_3	*true*	*false*	*false*	*false*	*false*	*false*	*false*
l_4	*false*	*false*	*false*	*false*	*false*	*false*	*true*
l_5	$W_{5,5}$	$W_{5,6}$	*false*	*false*	*false*	*false*	*false*
l_9	*true*	*false*	*false*	*false*	*false*	*false*	*false*
t_0	*false*	*false*	*false*	*false*	*false*	*false*	*true*
t_5	*false*	*false*	*false*	*true*	*false*	*false*	*false*
t_6	*false*	*false*	*false*	*false*	$W_{6,4}$	$W_{6,5}$	*true*
t_7	*false*	*false*	*true*	*false*	*false*	*false*	*false*
t_1	*false*	*false*	*false*	*true*	*false*	*false*	*false*

$$\vee(l_3, l_4, l_5, l_9, t_0, \wedge(t_1, t_5), t_6, t_7)>,$$

where:

- $W_{5,5}$ = "There are data mining tool that are not applied",
- $W_{5,6}$ = "The data mining tool is extracted",
- $W_{6,4}$ = "The data mining tool is rejected",
- $W_{6,5}$ = "The data mining tools are chosen".

The α-tokens obtain characteristic respectively *"Chosen data mining tools, criteria for the restricting data mining tool"* in place l_3, and *"Rejected data mining tool"* in place l_4.

The β_1-token that enters transition Z_2 (from place t_0) will unite with the original β_0-token, token that stays in place t_6.

The β_3-token that enters place t_5 obtains characteristic *"Selected data mining tools"*.

The β-tokens that enter places t_2 (from place t_7) and t_3 (from place t_5) do not obtain new characteristics.

$$Z_3 = <\{l_6, l_7, l_{11}\}, \{l_7, l_8\},$$

	l_7	l_8
l_6	*true*	*false*
l_7	$W_{7,7}$	$W_{7,8}$
l_{11}	*true*	*false*

$$, \vee(l_6, l_7, l_{11})>,$$

where:

- $W_{7,7}$ = "There is a next step in the current data mining technique",
- $W_{7,8}$ = "The next step in the current data mining technique is chosen".

The α-token that enters place l_7 do not obtains new characteristic and obtain characteristic *"Current step in the data mining technique"* in place l_8.

$$Z_4 = <\{t_3, t_8\}, \{t_7, t_8\}, \begin{array}{c|cc} & t_7 & t_8 \\ \hline t_3 & false & true \\ t_8 & W_{8,7} & W_{8,8} \end{array}>,$$

where:

- $W_{8,7}$ = *"The data mining tools for the goal are chosen"*,
- $W_{8,8} = \neg W_{8,7}$.

The β-tokens that enter place t_8 do not obtain new characteristic and obtain characteristic *"Goal, chosen data mining tools"* in place t_7.

$$Z_5 = <\{l_8\}, \{l_9, l_{10}, l_{11}\}, \begin{array}{c|ccc} & l_9 & l_{10} & l_{11} \\ \hline l_8 & W_{8,9} & W_{8,10} & W_{8,11} \end{array}>,$$

where:

- $W_{8,9}$ = *"The next data mining tool have to be applied"*,
- $W_{8,10}$ = *"The last data mining tool is applied"*,
- $W_{8,11}$ = *"The next step in the current data mining technique have to be applied"*.

The α-tokens that enter places l_9 and l_{11} do not obtain new characteristic and obtain characteristic *"Goal, data mining tools, evaluation of the data mining tool"* in place l_{10}.

3 Conclusion

The research presented in this paper is a continuation of previous investigations into the modelling of information flow with a typical university. The framework in which this is done is the theory of Generalized Nets (GNs) (and sub-GNs where appropriate).

References

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

[2] Atanassov, K. Generalized Nets, World Scientific. Singapore, New Jersey, London, 1991

[3] Chattamvelli, Data Mining Methods, Narosa Book Distributors, Pvt, Ltd, 2008.

[4] Ian Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, ISBN 0120884070, 2005.

[5] K. Cios, W. Pedrycz, R. Swiniarski, L. Kurgan, Data Mining: A Knowledge Discovery Approach, Springer, ISBN: 978-0-387-33333-5, 2007.

[6] Mark Hornick, Erik Marcade, Sunil Venkayala, Java data mining: strategy, standard, and practice, A practical Guide for Architecture, Design, and Implementation, 2006.

[7] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth, CRISP-DM 1.0, Step-by-step data mining guide, SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands), 2000.

[8] Sumathi, S., S.N. Sivanandam, Introduction to Data Mining Principles and its Applications, Studies in Computational Intelligence, Springer, Vol. 29, 2006.

[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. Foundations and Applications, 2009 (in press)

[10] Winston, P. Artificial Intelligence. Reading MA, Addison-Wesley, 1977.