Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics. Volume I: Foundations

Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics Volume II: Applications

Editors

Krassimir T. Atanassov
Michał Baczyński
Józef Drewniak
Janusz Kacprzyk
Maciej Krawczak
Eulalia Szmidt
Marcin Wygralak
Sławomir Zadrożny

Editors

Krassimir T. Atanassov
Władysław Homenda
Olgierd Hryniewicz
Janusz Kacprzyk
Maciej Krawczak
Zbigniew Nahorski
Eulalia Szmidt
Sławomir Zadrożny

SRI PAS IBS PAN

Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics
Volume II: Applications

Systems Research Institute
Polish Academy of Sciences

Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics
Volume II: Applications

Editors

Krassimir T. Atanassov
Władysław Homenda
Olgierd Hryniewicz
Janusz Kacprzyk
Maciej Krawczak
Zbigniew Nahorski
Eulalia Szmidt
Sławomir Zadrożny

IBS PAN  **SRI PAS**

Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
www.ibspan.waw.pl

ISBN 9788389475305

Generalized net model of the phases of the data mining process

E. Sotirova[1] and D. Orozova[2]

[1]Prof. Asen Zlatarov University, Bourgas-8000, Bulgaria,
esotirova@btu.bg
[2]Free University of Bourgas, Bourgas-8000, Bulgaria,
orozova@bfu.bg

Abstract

In the paper is constructed a Generalized Net (GN) model 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. The opportunity of using GNs as a tool for modeling such processes is analyzed as well.

Keywords: data mining, e-learning, generalized nets.

1 Main tools for data mining

Data mining is the process of discovery of hidden patterns and dependencies in data.

Main data mining techniques are Advanced analysis, Predictive analysis, Comparative analysis, Cluster analysis, Decision trees, Association rules, Neural network. These rules might be generated by a process of discovery and research of combinations of rules or be derived from the solution trees. Also neural networks and genetic algorithms might be used. The model utilizes repetitive evolutionary models, while fitness function is used for the selection of a particular feature or deviation.

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.

The process of data mining consists of the following phases:

Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics. Volume II: Applications *(K.T.Atanassow, W.Homenda, O.Hryniewicz, J.Kacprzyk, M.Krawczak, Z.Nahorski, E. Szmidt, S. Zadrożny, Eds.), IBS PAN - SRI PAS, Warsaw, 2009.*

1. ***Business understanding*** is the initial phase, focusing on the definition of the research objectives within the problem area and the respective user defined requirements. At the end of this phase, this knowledge has to be converted into definitions of data mining problems, and a preliminary plan designed to achieve the objectives.
2. ***Data understanding*** starts with the initial data collection and continues with activities aimed at intensifying the researcher's comprehension of the nature of these data. At this phase it is necessary to identify problems, related to the quality of the data, to develop the primary perception of the character of the data, as well as to figure out the remarkable data subsets, in order to shape up the initial hypotheses about the information hidden within the data.
3. ***Data preparation*** covers all aspects of the transformation of initial "raw" data into the final data set (i.e. the data that will be used by the modelling tools). Often, the data preparation phase needs to be performed repeatedly at different moments of time. The tasks within this phase include data table, case, and attribute selection as well as transformation and cleaning of data.
4. ***Modelling*** consists of selection and application of various modelling techniques, aiming at the extraction of regular patterns from the data. The parameters of the models are calibrated to their optimal values. Since some models possess their specific requirements of the data form, stepping back to the data preparation phase is often needed.
5. ***Model evaluation*** represents the careful review of all steps, executed to construct the particular model, in order to confirm its capability of achieving the objectives. At the end of the phase, a decision on the use of the data mining results is reached.
6. ***Deployment*** is related to the need of supervision and utilization strategy. For instance, at this phase a decision may be taken when and how the model may be repeated and under what circumstances.

Herewith, a model has been constructed with the data mining research tools of the generalized nets - CRISP-DM (Cross-Industry Standard Process Model of Data Mining). Figure 1 illustrates the CRISP-DM process.

The process outlines the tasks, related to every phase of the data mining process, as well as the relationships between the phases and the tasks. As with most standards, CRISP-DM does not cover all phase-task relationships, since it depends on the project objectives, user experience and needs, as well as the data specifics.

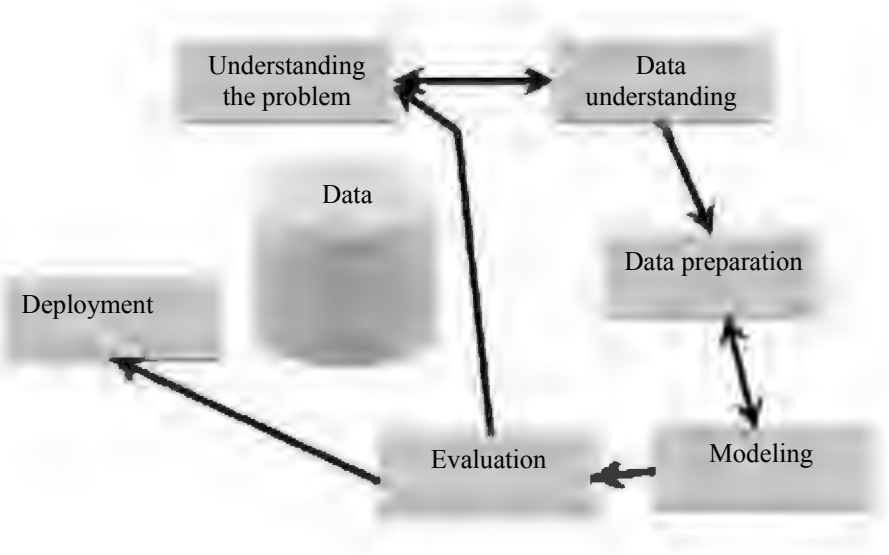


Figure 1: Scheme of the Data Mining process, CRISP-DM (Cross-Industry Standard Process Model of Data Mining).

2 Generalized net model of CRISP-DM

The GN contains the following set of transitions:

$$A = \{ Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8 \},$$

and they 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.

The forms of the transitions are the following.

α_1, α_2, α_3-token enter the net from places l_1, l_2 and l_3 with initial characteristics respectively:

"business objectives" in place l_1,
"users requirements" in place l_2,
and "resources" in place l_3.

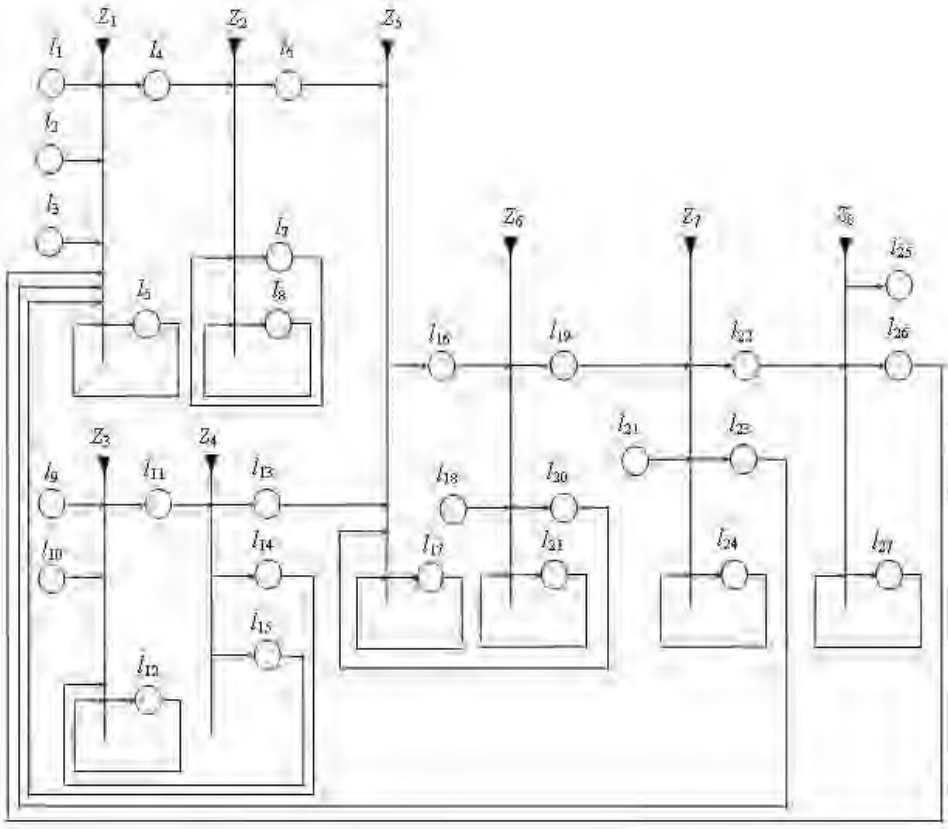


Figure 2: Generalized net model of the CRISP-DM

Let w is the token in the generalized net and it can splits, the new tokens will be noted by w', w'' and so on.

$$Z_1 = <\{l_1, l_2, l_3, l_5, l_{14}, l_{23}, l_{26}\}, \{l_4, l_5\},$$

	l_4	l_5
l_1	*false*	*true*
l_2	*false*	*true*
l_3	*false*	*true*
l_5	$W_{5,4}$	$W_{5,5}$
l_{14}	*false*	*true*
l_{23}	*false*	*true*
l_{26}	*false*	*true*

$>$,

where:

$W_{5,4}$ = “a data mining problem is defined”,
$W_{5,5} = \neg W_{5,4}$.

The α_1, α_2 and α_3-token that enter place l_5 do not obtain new characteristics. They merge in a β-token with characteristic

“data mining problem”.

When the $W_{5,4}$ is true the β-token enters place l_4 and it do not obtains new characteristic.

$$Z_2 = <\{l_4, l_7, l_8\}, \{l_6, l_7, l_8\},$$

	l_6	l_7	l_8
l_4	*false*	*false*	*true*
l_7	$W_{7,6}$	$W_{7,7}$	*false*
l_8	*false*	$W_{8,7}$	$W_{8,8}$

$>$,

where:
$W_{7,6}$ = “a preliminary plan is ready”,
$W_{7,7} = \neg W_{7,6}$,
$W_{8,7}$ = “definitions of data mining tasks are ready”,
$W_{8,8} = \neg W_{8,7}$.

On the first activation of the transition Z_2 the β-token (from place l_4) enter place l_8. On the next activation of the transition Z_2 (when the $W_{8,7}$ is true) this β-token enter place l_7 with characteristic

“data mining problem , definitions of data mining tasks”.

On the next activation of the transition Z_2 (when the $W_{7,6}$ is true) the β-token enter place l_6 with characteristic

“data mining problem , definitions of data mining tasks,
preliminary plan to achieve the objectives”.

α_4, α_5-token enter the net from places l_9 and l_{10} with initial characteristics respectively:

“initial data collection” in place l_9,
and “criterion for evaluation of the initial data collection” in place l_{10}.

$$Z_3 = <\{l_9, l_{10}, l_{12}, l_{15}\}, \{l_{11}, l_{12}\},$$

	l_{11}	l_{12}
l_9	*false*	*true*
l_{10}	*false*	*true*
l_{12}	$W_{12,11}$	$W_{12,12}$
l_{15}	*false*	*true*

$>$,

where:
$W_{12,11}$ = “a primary perception of the character of the data is developed”,
$W_{12,12} = \neg W_{12,11}$.

The α_4 and α_5-token that enter place l_{12} do not obtain new characteristics. They merge in a γ-token with characteristic

"primary perception of the character of the data".

When the $W_{12,11}$ is true the γ-token enters place l_{11} and it do not obtains new characteristic.

$$Z_4 = <\{l_{11}\}, \{l_{13}, l_{14}, l_{15}\}, \begin{array}{c|ccc} & l_{13} & l_{14} & l_{15} \\ \hline l_{11} & W_{11,13} & W_{11,14} & W_{11,15} \end{array} >,$$

where:

$W_{11,13}$ = "the data is obtained",

$W_{11,14}$ = "the predefining the data mining problem is necessary",

$W_{11,15}$ = "the cleaning of data or obtaining new data is necessary".

The γ-tokens that enter places l_{13}, l_{14} and l_{15} obtain characteristics respectively:

"necessary data mining problem for data",
"data mining problem for predefinition",
and "data for cleaning or request for new data".

$$Z_5 = <\{l_6, l_{13}, l_{17}, l_{20}\}, \{l_{16}, l_{17}\}, \begin{array}{c|cc} & l_{16} & l_{17} \\ \hline l_6 & false & true \\ l_{13} & false & true \\ l_{17} & W_{17,16} & W_{17,17} \\ l_{20} & false & true \end{array} >,$$

where:

$W_{17,16}$ = "the initial "raw" data were transformed to the final data set",

$W_{17,17} = \neg W_{17,16}$.

The β and γ-token that enter place l_{17} (from places l_6 and l_{13}) do not obtain new characteristics. They merge in a δ-token with characteristic

"data mining problem , definitions of data mining tasks,

preliminary plan to achieve the objectives, data that will be used by the modelling tools".

When the $W_{17,16}$ is true the δ-token enters place l_{16} and it do not obtains new characteristic.

At this phase, the user provides the settings of the mining functions, and optionally in case of larger control needed, may choose the algorithms for model construction, as well as their specific settings.

α_6-token enters the net from place l_{18} with initial characteristic

"modelling techniques".

$$Z_6 = <\{l_{16}, l_{18}, l_{21}\}, \{l_{19}, l_{20}, l_{21}\},$$

	l_{19}	l_{20}	l_{21}
l_{16}	*false*	*false*	*true*
l_{18}	*false*	*false*	*true*
l_{21}	$W_{21,19}$	$W_{21,20}$	$W_{21,21}$

>,

where:
$W_{21,19}$ = "the data model is constructed",
$W_{21,20}$ = "it is necessary stepping back to the data preparation phase",
$W_{21,21} = \neg W_{21,19}$ & $\neg W_{21,20}$.

The δ- and α_6-token that enter place l_{21} (from places l_{16} and l_{18}) do not obtain new characteristics. They merge in a χ-token with characteristic

"data mining problem, data model".

Although the models constructed in the modelling phase may exhibit high quality, they may not properly match the business requirements yet. At this phase it might prove necessary to return to previous steps in order to determine if an important aspect of the problem has been missed. The aim of this phase is to determine whether the model may be utilized in a business application or a business process.

α_7-token enters the net from place l_{21} with initial characteristic "the test data".

$$Z_7 = <\{l_{19}, l_{21}, l_{24}\}, \{l_{22}, l_{23}, l_{24}\},$$

	l_{22}	l_{23}	l_{24}
l_{19}	*false*	*false*	*true*
l_{21}	*false*	*false*	*true*
l_{24}	$W_{24,22}$	$W_{24,23}$	$W_{24,24}$

>,

where:
$W_{24,22}$ = "the model may be utilized in a business application or a business process",
$W_{24,23} = \neg W_{24,22}$,
$W_{24,24}$ = "the model evaluation is not finished".

The χ- and α_7-token that enter place l_{24} (from places do not obtain new characteristics. They merge in a ε-token with characteristic

"evaluated data mining problem".

The ε-token splits in two new tokens that enter places l_{22} and l_{23} with characteristic respectively:

"the data mining model for a business application or a business process"
and "data mining model, the aspect of the problem has been missed".

The deployment phase in CRISP-DM concentrates on the packaging of the results from the data mining process, i.e. the knowledge extracted from the data, and acquisition of experience for specific business problems for users.

$$Z_8 = <\{ l_{22}, l_{27} \}, \{l_{25}, l_{26}, l_{27}\},$$

	l_{25}	l_{26}	l_{27}
l_{22}	*false*	*false*	*true*
l_{27}	$W_{27,25}$	$W_{27,26}$	$W_{27,27}$

$>$,

where:
$W_{27,25}$ = “the model is deployed”,
$W_{27,26}$ = “the model has to be changed”,
$W_{27,27}$ = “the model is in the deployment phase”.

The ε-tokens that enter places l_{25} and l_{26} obtain characteristics respectively:
“deployed model”,
and “model, conditions for model rebuilding”.

3 Exemplary realization

We shall utilize the tools of Java DM [2], which provides a framework of data mining functions. The applications programming interface (API) of JDM allows us to determine how to implement the solutions by the use of settings and data objects. We pass consecutively through the all the phases:

3.1 Problem definition

This stage comprises of determining whether a client matches the criteria of a new affinity card program.

3.2 Data understanding

All data are contained in the main table Sales History. It contains demographic data, record of purchase, as wells as information about previous card programs. The program manipulates with user data, described by the relational model.

3.3 Data preparation

The help tables, generated by the program, are:

- MINING_DATA_BUILD_V – contains data about the client: demographic data, record of purchase and responses to offers from previous promotions. This information is necessary for the model construction.

- MINING_DATA_TEST_V – contains data about former clients: demographic data and responses to offers from previous promotions. This information is needed for testing the model.
- MINING_DATA_APPLY_V – collects data about future clients: demographic data and records of purchase. This information is needed for predictions of the clients' responses to the new card program.

Now is the moment to determine the attributes that will be used for the data mining, namely the "Age" and "Average annual income" attributes. If the number of teachers is a way too big, subsets of teachers may be introduced here.

3.4 Modelling by means of a decision tree

Decision tree is one of the most popular algorithms due to the easy comprehension of the way it makes classifications and prognoses. A decision tree is a dendroid structure for which in each node an analytic test of the attribute value is made, and every branch represents the outcome of the test, leaves being classes or class distributions. The decision tree produces rules, and in this way the user understands not only how and why a prediction is made, but these rules are further useful for the segmentation of the set. The decision tree algorithm is widely used in practice for classification purposes and in some applications it is implemented to support regression as well.

In the example, the root represents all n-teachers in the table. With respect to the first attribute, "Age", the algorithm learns that those aged less than 46 years are probable leavers. Thus, the first node distributes the data to nodes 1 and 2 on the basis of the teachers' age. Then, each of these nodes distributes the data on the basis on the condition related to the "Average annual income" attribute.

Each tree node is associated with a rule that predicts the target value with certain degrees of **"Confidence"** and **"Support"**.

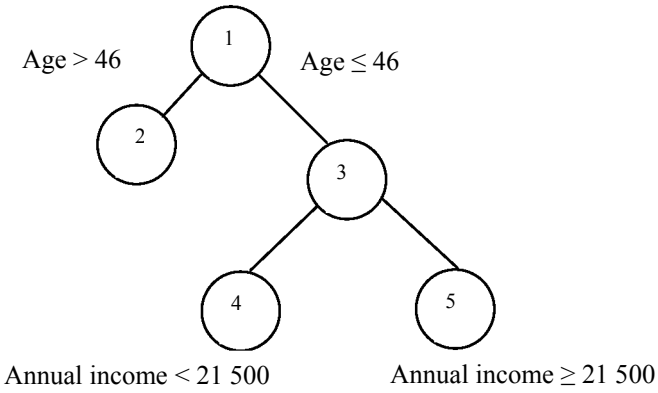
- "Confidence" is a measure of the probability with which the tree root predicts the target value. This is the proportion between the number of cases in the node, for which the condition has been correctly predicted, and the total number of cases in the node.
- "Support" is the proportion between the number of cases in the node, for which the condition has been correctly predicted, and the total number of cases in the table. The measure here is the number of cases in the database, which are related to the condition.

Let us choose 10 random teachers from the TEACHERS_DATA table, so that half of them have corresponding value "true" for the "Satisfaction" target attribute.

ID of teacher	Age	Length of service	Average annual income	Satisfaction
1	51	4,5	11 500	+
2	45	15,0	3 500	-
3	36	3,4	21 500	+
4	47	6,1	36 000	+
5	42	14,5	7 000	-
6	50	2,5	15 000	+
7	40	11,9	6 000	-
8	31	4,10	2 000	-
9	38	10,2	5 500	-
10	37	7,5	31 500	+

Using the data from the table, the algorithm constructs a decision tree, learning about the first attribute of "Age" that those teachers, who are more than 46 years old, are satisfied. The algorithm finds the maximal age, for which the target attribute of "Satisfaction" has negative value "false", i.e. the teacher is a probable leaver, and this maximal age is max(Age) = 45.

Thus the first node distributes the data in nodes 2 and 3 on the basis of the teachers' age. Then, node 3 divides the data in nodes 4 and 5 on the basis of the condition related to the "Average annual income" attribute.



There follows a reference for each node of the tree: the rule, associated with it, the value of the target function, the number of cases in the node and the measures of "confidence" and "support".

Node	Rule	Target value	Number of cases	Confidence	Support
1		Satisfied	10	5/10=0.5	5/10=0.5
2	Age > 46	Satisfied	3	3/3 = 1	3/10=0.3
3	Age ≤ 46	Unsatisfied	7	5/7=0.7	7/10=0.7
4	Age ≤ 46 & Income < 21 500	Unsatisfied	5	5/5 = 1	5/10=0.5
5	Age ≤ 46 & Income ≥ 21 500	Satisfied	2	2/2 = 1	2/10 =0.2

JDM API also works on this principle when constructing the model with the decision tree algorithm. Thus, using the method buildModel(), a tree is constructed with JDM, determining the rules for distribution of the available data in the database, retrieving detailed information about the decision tree and the rules, which the model construction is based on.

JDM permits the specification of the possible expenses, related to the incorrect predictions a within a weight matrix. The weight matrix is an N×N table that determines the expenses, related to incorrect predictions, where N is the number of possible target values. In JDM, the weight matrix is determined as an object that may be set up separately by the user. For a more precise model setup the class TreeHomogeneityMetric is also used. This class, defined in JDM API is used for "trimming" the tree, i.e. it introduces constraints that help avoiding the infinite tree depth or empty nodes. For instance, the following contstraints are introduced:

Maximal depth: 7
MinNodeSizeForSplit (percentage): 0,1
MinNodeSizeForSplit (count): 20
Min Node Size (count): 10
Min Node Size (percentage): 0,05

3.5 Evaluation stage

It is important to evaluate the quality of the model, before it is utilized for prediction of real data. The data are of two types: one needed for building the model, and another needed for its testing. The test data are those records, which have not been utilized for the model construction, aimed at giving a correct and precise assessment of the model with respect to the accuracy of the predictions it makes. JDM maintains four tests for classification models: prediction accuracy, chaos matrix, receiver operating characteristics (ROC); lift. The testing is executed by the method computeTestMetrics().

Once the model efficiency is calculated using test data, the user makes the decision for applying the model to real data. Some algorithms may utilize a part of the input attributes as obligatory for the application of the final model. This is called *model signature*.

The model that constitutes the program also has its own signature, which may be seen in the output data of the program execution.

Details of the model signature: (ATTRIBUTE-NAME, Attribute type)
(NAME, stringType)
(AGE, doubleType)
(GRADE, stringType)
(ACADEMIC-RANK, stringType)
(EDUCATION, stringType)
(INCOME, doubleType)
(PRACTICE, doubleType)
(GENDER, stringType)
(SATISFACTION, yes/no)

After the end of the program execution over full real data, we have a constructed and tested model. Complete information about the model is shown (root ID, total tree depth, total number of nodes and leaves) and each of the retrieved rules corresponds to a node in the tree, labeled with the respective number. Thus, we already know that the algorithm has placed in node 9 all teachers who are older than 46 years and obtain average annual income larger than 21 500.

On the basis of this information, we can graphically represent the decision tree, as shown on Figure 4.

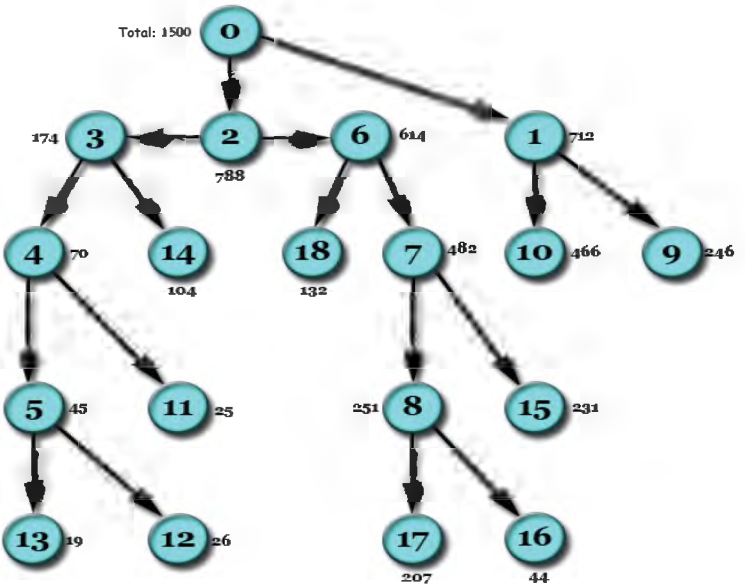


Figure 4: Graphic representation of the decision tree, generated by the system

3.6 Integration and deployment

When we dispose of the ready tested model, we can apply it to real data, in order to check its functionality.

4. Conclusion

The research expounded 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). While the order of procedure might vary from one institution to another, the processes are almost invariant, so that the development of the GN in this paper can be readily adapted or amended to suit particular circumstances, since each transition is constructed in a transparent manner.

References

[1] 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.

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

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

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

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

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

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

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

The papers presented in this Volume 2 constitute a collection of contributions, both of a foundational and applied type, by both well-known experts and young researchers in various fields of broadly perceived intelligent systems.
It may be viewed as a result of fruitful discussions held during the Eighth International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets (IWIFSGN-2009) organized in Warsaw on October 16, 2009 by the Systems Research Institute, Polish Academy of Sciences, in Warsaw, Poland, Centre for Biomedical Engineering, Bulgarian Academy of Sciences in Sofia, Bulgaria, and WIT – Warsaw School of Information Technology in Warsaw, Poland, and co-organized by: the Matej Bel University, Banska Bistrica, Slovakia, Universidad Publica de Navarra, Pamplona, Spain, Universidade de Tras-Os-Montes e Alto Douro, Vila Real, Portugal, and the University of Westminster, Harrow, UK:

http://www.ibspan.waw.pl/ifs2009

The Eighth International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets (IWIFSGN-2009) has been meant to commence a new series of scientific events primarily focused on new developments in foundations and applications of intuitionistic fuzzy sets and generalized nets pioneered by Professor Krassimir T. Atanassov. Moreover, other topics related to broadly perceived representation and processing of uncertain and imprecise information and intelligent systems are discussed.

We hope that a collection of main contributions presented at the Workshop, completed with many papers by leading experts who have not been able to participate, will provide a source of much needed information on recent trends in the topics considered.

ISBN-13 9788389475305
ISBN 838947530-8