Advanced Methods for Knowledge Extracting for Complex Processes

Bulgarian Academy of Sciences

Institute of Information and Communication Technologies

Boriana Vatchova M.Sc, PhD

e-mail: *boriana.vatchova@gmail.com*

International Conference AUTOMATICS AND INFORMATICS'2013
Bulgaria, Sofia, October 3 - 7, 2013

OUTLINE

1. Introduction
2. Essence of the Method of MLPF
3. Models of complex processes under uncertainty using MLPF
   - 3.1 Production Rule Model
   - 3.2. Network Model
4. Conclusion

Introduction

Definition for Data Mining

‘Non-trivial extraction of implicit, previously unknown, and potentially useful information from data’ W.Frawley, G.Piatetsky –Shapiro and S. Mathews

Data mining is also known as knowledge discovery in databases.

Introduction

Data mining can also be considered as coherent merging of information from multiple sources.

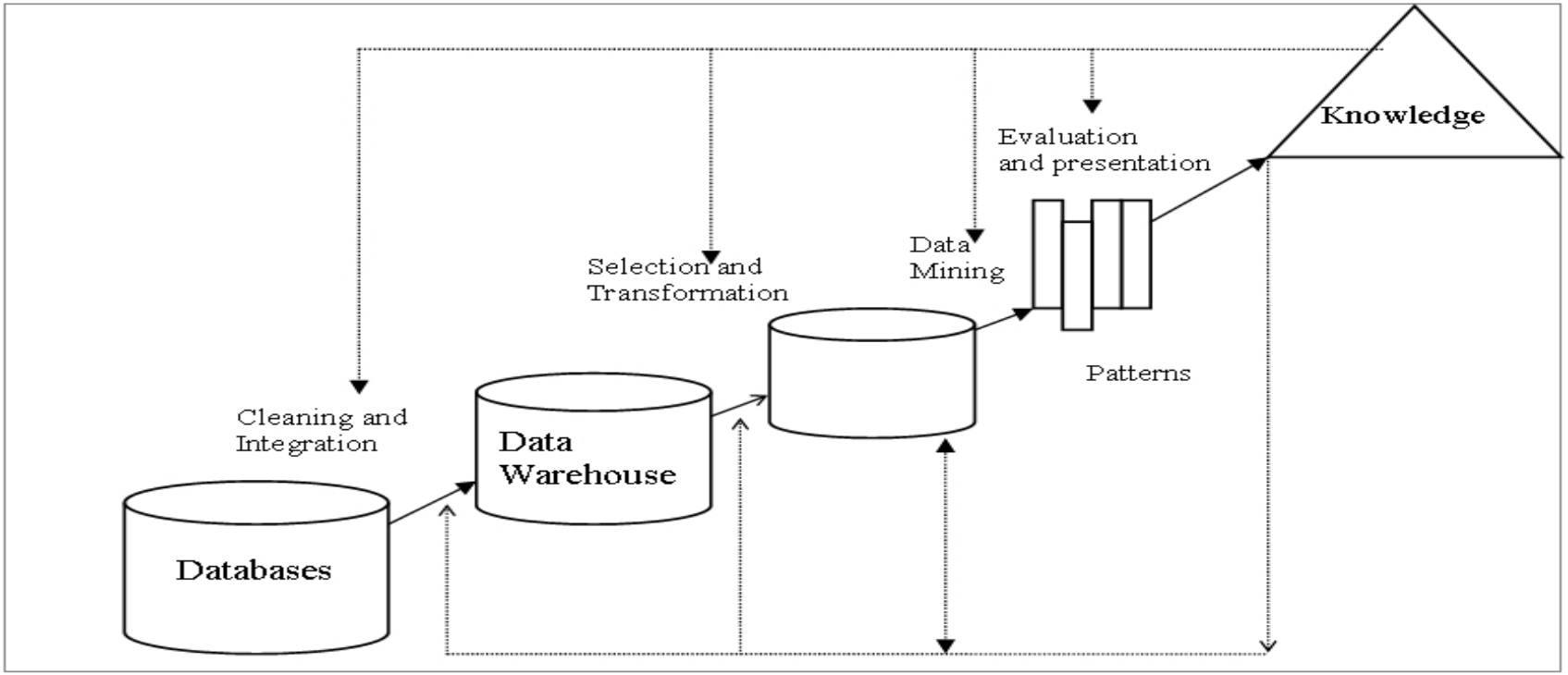

Fig.1 Data mining as a step in the process of knowledge discovery

Introduction

Existing methods for knowledge presentation in intelligent systems

- **Logical formulas**

Logical models for knowledge presenting are based on conception of formal system. One formal system is presented with next four elements M =<T,P,A,F> where T is set of base elements, P is set of syntaxes rules, A is set of true axioms and P is set of rules for inferences.

- **Production rules**

Production models are combination of production rules which are related each others in form of the following type:

If <antecedent> Then < consequent >

Introduction

- **Semantic Nets**

Semantic nets are other type for knowledge presenting. They are familiar as their graphical visualisation in form of nodes and vertices.

- **Frames**

Frames or list of facts are the next form for knowledge presenting in intelligent systems. They are used when investigated data are characters or words.

Introduction

Existing methods for knowledge extraction:

- **Expert oriented method**

The methods of expert's opinion are based on expert's experience in particular research field.

- **Data mining methods**

Data Mining methods are based on data grouping according to similar properties like clusterisation, classification and grouping.

Introduction

Disadvantages of existing methods for knowledge extraction:

- Subjective and intuitive (expert oriented methods)
- Not suitable for real time (data mining methods)
- Not suitable for the complex processes (data mining methods)

Introduction

Attributes of complex processes

Quantitative complexity

- Large number of inputs
- Large number of state parameters
- Large number of outputs

Qualitative complexity

- Non-linearity, non-stationary and uncertainty
- Environmental disturbances
- Immeasurable of some inputs

Introduction

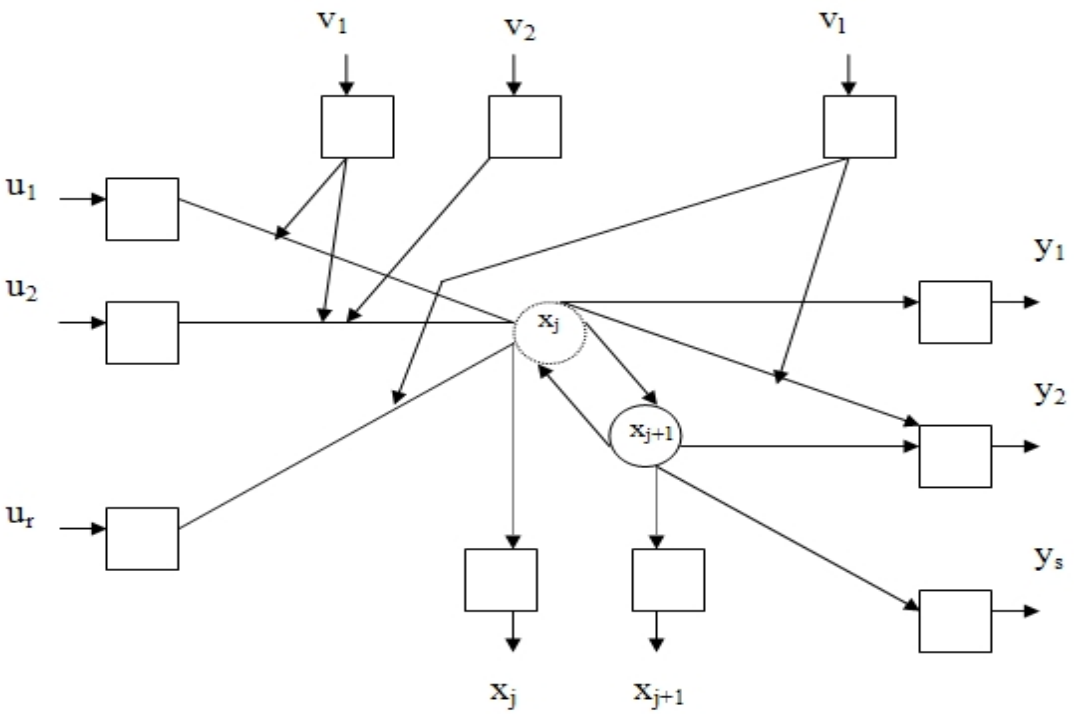

Fig.2 Structure of model of complex process

Introduction

- U (u_1,u_2,…,u_r)- controllable inputs;
- V(v_1,v_2,…,v_l)- parameters of the environment disturbances;
- X(x_1,x_2,…,x_d)- states parameters;
- Y(y_1,y_2,…,y_s)- outputs;

Introduction

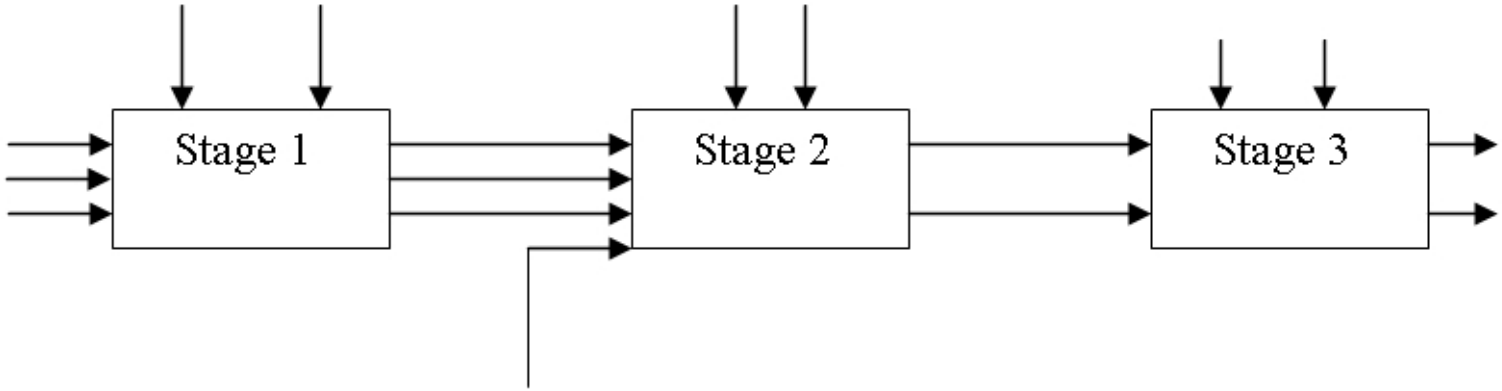

Fig.3 Multi-stage system with sequence composition

Introduction

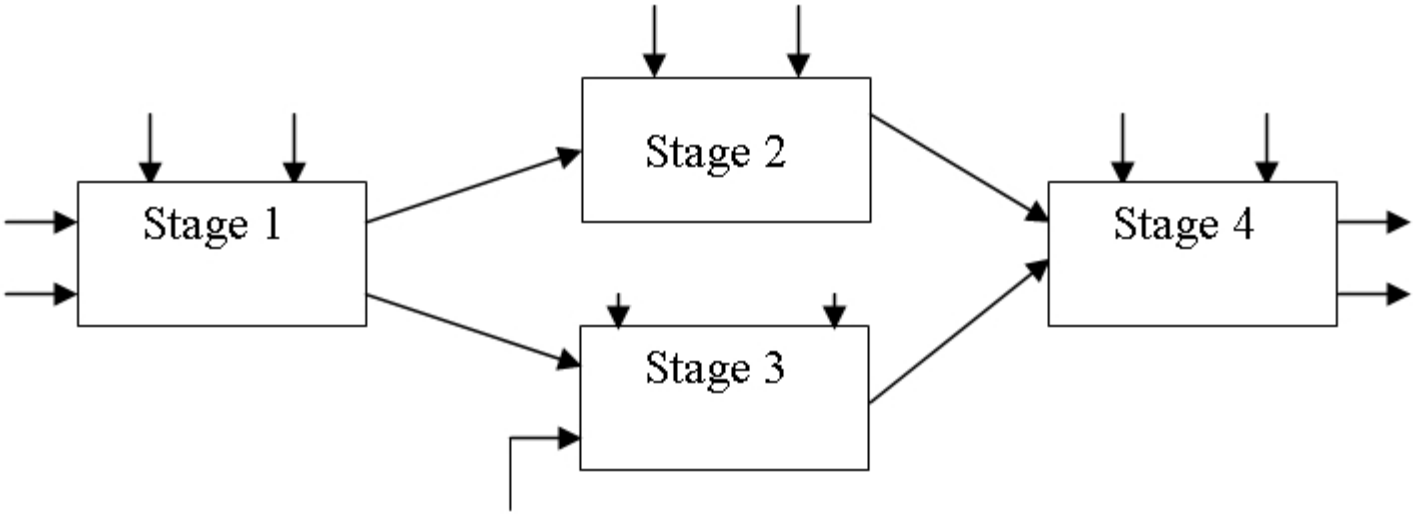

Fig.4 Multi-stage system with parallel composition

Introduction

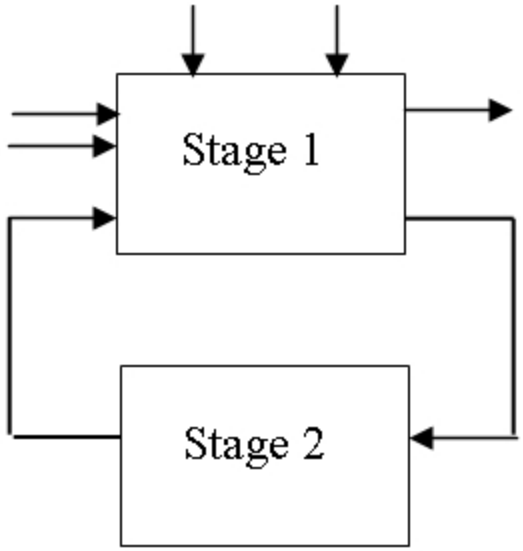

Fig.5 Multi-stage system with feedback

Introduction

Modeling problems for complex processes:

- The high accuracy of the model of the complex process is incompatible with increasing of complexity of the object (L. Zadeh)

- Low efficiency of process model

Essence of the Method of MLPF

The models of the processes are:

- knowledge bases (KB) of production rules, which include probability of occurrences [7]

- updatable multi-layer network structure [8]

Essence of the Method of MLPF

Novel approach for knowledge extraction MLPF [6,11]:

- *multi-valued logical and probabilistic functions*

$$<Ly, p\{Ly\}> = F(Lx_1, Lx_2, \ldots, Lx_n) \quad (1)$$

where

$Lx\{L_i\}$, $i=1 \div n$ - a set of logical values of the arguments;

Ly_j, $j=1 \div m$ – set of logical values of the outputs;

$$Ly_{eq}=F_1\{GLN_r\}, \qquad GLN_r(\tau)=F_2\{Lx_{ij}, W\} \quad (2)$$

$$p\{Ly_{eq}\}=P_1\{GLNr\} \qquad p\{GLN_r\}=P_2\{Lx_{ij}, W\} \quad (3)$$

Essence of the Method of MLPF

Table 1 MLPF for three degree logical system [8,11]

Lx_1		a_1	a_1	a_1	a_1	a_1	a_1	a_1	a_2	a_2	.....	.....	a_3
Lx_2		a_1	a_2	a_2	a_2	a_3	a_3	a_3	a_1	a_2	.....	.....	a_3
Lx_3		a_1	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_1	.....	.....	a_3
Ly_1	a_1	p_{1111}	p_{abcd}										
	a_2	p_{1112}		p_{1222}									
	a_3	p_{1113}											P_{3333}
Ly_2	a_1	p_{1111}							p_{abcd}				
	a_2	p_{1112}					p_{1322}						
	a_3	p_{1113}						p_{1333}					P_{3333}

Essence of the Method of MLPF

where Lx_1, Lx_2, Lx_3 are inputs and Ly_1 and Ly_2 are the outputs;

a_1,a_2 and a_3 are logical values with a meaning for small, medium and large;

Pabcd is probability (frequency) of occurrences for the three inputs and one output;

W(t) no apparent argument (factor) for inputs which is including in frequency of occurrence Pabcd

Models of complex processes under uncertainty using MLPF

3.1. Production Rule Model [6,7]

Using new data sets in real time creates packages of numerical values for inputs and outputs, which are updated values of MLPFs [7] . The model or the updated knowledge base is a combination of production rules with the following structure:

If < *logical values of measurable inputs*> **Then** < *logical values of the outputs supplemented with a probability of occurrences*>

or

If < Lx_1, Lx_2,Lx_3 > **Then** < Ly_1, Ly_2 >

Models of complex processes under uncertainty using MLPF

- The process with three inputs with logical values and three outputs is presented. The experimental data sets are 56 and part of them are shown in Table 7, [6].
- The flotation process from the mining industry is a typical complex process that deals with the enrichment of raw ore.
- Flotation is implemented by processing a mixture of finely ground ore, water and reagents called pulp.

Models of complex processes under uncertainty using MLPF

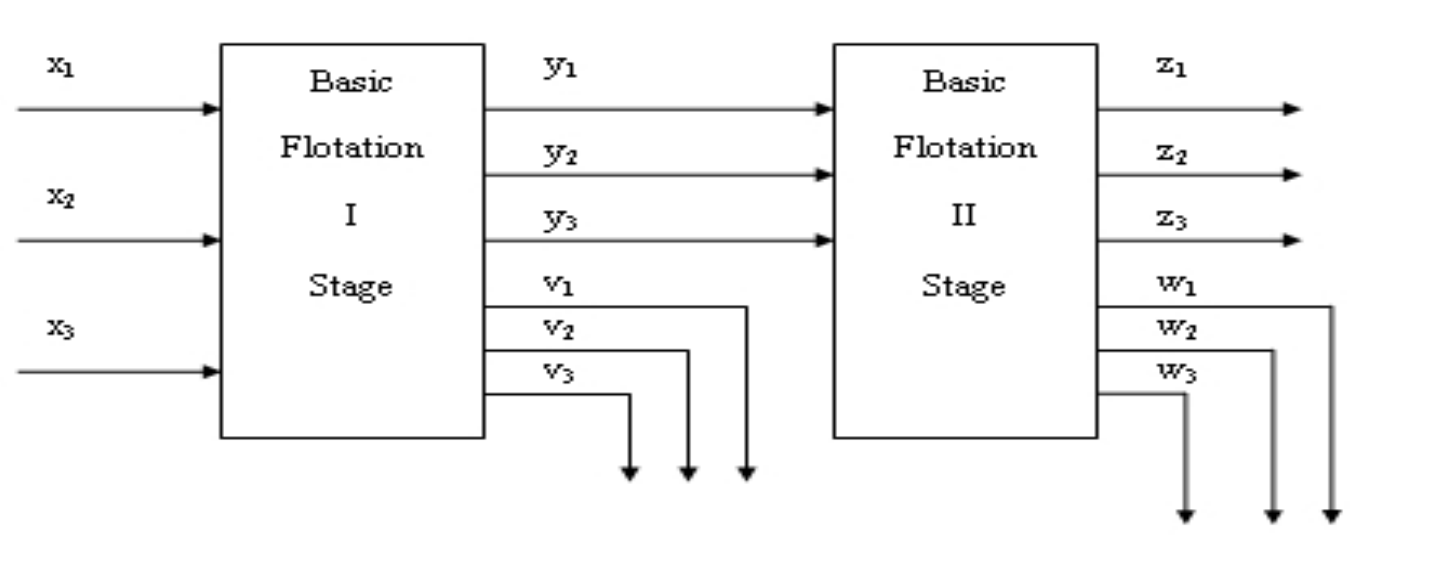

Fig. 6 General structure of basic flotation in two stages

Models of complex processes under uncertainty using MLPF

Inputs are controllable and measurable parameters of the pulp [6].

x_1 – percentage of copper content in incoming pulp, stage I ;

x_2 – percentage of iron content incoming pulp, stage I;

x_3 – capacity of incoming pulp, stage I [m^3/h].

y_1, y_2, y_3–analogical parameters of the output concentrate in stage I, incoming for flotation in stage II (outputs for stage I and inputs for stage II);

v_1, v_2, v_3 – analogical parameters of the waste of the pulp used for extraction of the useful components (waste for stage I);

z_1, z_2, z_3 – analogical parameters of the concentrate in stage II (output for stage II);

w_1, w_2 , w_3 – analogical parameters of the waste of the pulp in stage II.

Models of complex processes under uncertainty using MLPF

A seven-degree logic system is perceived where the inputs and the outputs have logical values : VVS, VS, S, M, L, VL, VVL, where VVS is ‘very very small’, VS is ‘very small’, S is ‘small’, M is ‘medium’, L is ‘large’, VL is ‘very large’, VVL is ‘very very large’ [6,11].

Models of complex processes under uncertainty using MLPF

№	x_1	x_2	x_3	y_1	y_2	y_3
1	0,406	2,891	16,02	5,142	13,90	12,10
2	0,391	2,89	18,28	0,964	7,856	39,95
3	0,4089	2,952	14,61	2,491	3,668	16,9
4	0,409	2,943	14,83	4,974	13,78	11,22
5	0,419	2,995	13,81	5,39	13,75	11,26
6	0,418	2,967	13,78	4,573	13,14	13,02
7	0,407	2,925	15,77	1,696	8,89	39,82
8	0,389	2,876	18,45	2,77	4,755	19,05
9	0,404	2,994	14,48	0	3,701	40,4
10	0,402	2,937	15,19	0,453	7,155	40,05
11	0,418	2,999	13,5	0	0	23,47
12	0,408	2,959	14,363	3,267	6,366	15,55
13	0,408	2,974	14,3	1,888	9,166	39,7
14	0,409	2,973	14,17	2,444	2,802	18,56
15	0,403	2,953	14,7	3,287	6,269	17,18
16	0,397	2,92	16,04	2,024	9,373	39,5
17	0,419	2,893	12,36	0,269	0	21,3
18	0,406	2,984	14,32	4,85	12,15	15,01
19	0,375	2,862	19,41	1,479	8,576	39,78
20	0,369	2,828	22,13	0,431	7,126	40,13
21	0,353	2,794	24,45	2,16	0,327	24
22	0,367	2,862	21,2	1,763	8,989	39,8
23	0,385	2,905	18,4	0,319	6,985	40,1

Table 7 Experimental data for the basic flotation process*

The purpose for knowledge extraction is to reveal relations for multiple repeating correspondences between logical values of inputs and outputs

*These data are part of full sets of inputs and outputs. The data are received under normal working conditions for a 24hour time interval

Models of complex processes under uncertainty using MLPF

GLN,		1	2	3	4	5
Z,		19	12	11	10	3
p{GLN,}		0,3455	0,2182	0,200	0,1818	0,0545
Lx_1		VVL	VVL	VVL	VL	VL
Lx_2		VVL	VVL	VVL	VVL	VVL
Lx_3		VL	M	L	VL	VVL
Ly_1	VVS	1	1	4	2	0
		0,053	0,083	0,364	0,2	0
	VS	3	2	1	0	0
		0,158	0,167	0,091	0	0
	S	2	4	0	2	2
		0,105	0,333	0	0,2	0,667
	M	0	2	2	0	0
		0	0,167	0,182	0	0
	L	2	0	0	0	0
		0,105	0	0	0	0
	VL	9	2	1	5	1
		0,474	0,167	0,091	0,5	0,333
	VVL	2	1	3	1	0
		0,105	0,083	0,273	0,1	0
Ly_2	VVS	2	1	2	2	1
		0,105	0,083	0,182	0,2	0,333
	VS	0	1	2	0	0
		0	0,083	0,182	0	0
	S	0	2	1	0	0
		0	0,167	0,091	0	0
	M	5	2	2	2	0
		0,263	0,167	0,182	0,2	0
	L	6	3	0	2	1
		0,316	0,125	0	0,2	0,333
	VL	4	1	3	3	1
		0,211	0,083	0,273	0,3	0,333
	VVL	2	2	1	1	0
		0,105	0,167	0,091	0,1	0
Ly_3	VVS	0	0	0	0	0
		0	0	0	0	0
	VS	0	1	1	0	0
		0	0,083	0,091	0	0
	S	5	4	3	6	1
		0,263	0,333	0,273	0,6	0,333
	M	9	1	3	1	0
		0,474	0,083	0,273	0,1	0
	L	1	1	1	1	1
		0,053	0,083	0,091	0,1	0,333
	VL	0	0	0	0	0
		0	0	0	0	0
	VVL	4	5	3	2	1
		0,211	0,417	0,273	0,2	0,333

Table 8 Multi-valued logical-probabilistic function for the flotation process for single delay between inputs and outputs.

Analogical results for multi-valued logical-probabilistic function by multiple (three, four and five) time delay between inputs and outputs .

Models of complex processes under uncertainty using MLPF

Table 8' MLPF for a limited number of data sets Lx_1, Lx_2, Lx_3 and Ly_1

GLN_r		1	2	3	4	5
Z_r		19	12	11	10	3
$p\{GLN_r\}$		0,3455	0,2182	0,200	0,1818	0,0545
Lx_1		VVL	VVL	VVL	VL	VL
Lx_2		VVL	VVL	VVL	VVL	VVL
Lx_3		VL	M	L	VL	VVL
Ly_1	VVS	1	1	4	2	0
		0,053	0,083	0,364	0,2	0
	VS	3	2	1	0	0
		0,158	0,167	0,091	0	0
	S	2	4	0	2	2
		0,105	0,333	0	0,2	0,667
	M	0	2	2	0	0
		0	0,167	0,182	0	0
	L	2	0	0	0	0
		0,105	0	0	0	0
	VL	9	2	1	5	1
		0,474	0,167	0,091	0,5	0,333
	VVL	2	1	3	1	0
		0,105	0,083	0,273	0,1	0

Models of complex processes under uncertainty using MLPF

GLN_i		1	2	3	4	5
Z_i		19	12	10	10	3
p{GLN}		0,3519	0,222	0,1852	0,1852	0,0556
Lx_1		VVL	VVL	VVL	VL	VL
Lx_2		VVL	VVL	VVL	VVL	VVL
Lx_3		VL	M	L	VL	VVL
Ly_1	VVS	0	3	2	1	2
		0	0,25	0,2	0,1	0,667
	VS	4	1	0	0	0
		0,211	0,083	0	0	0
	S	3	4	1	1	1
		0,158	0,333	0,1	0,1	0,333
	M	0	2	2	0	0
		0	0,167	0,2	0	0
	L	1	0	0	1	0
		0,053	0	0	0,1	0
	VL	8	0	4	6	0
		0,421	0	0,4	0,6	0
	VVL	3	2	1	1	0
		0,158	0,167	0,1	0,1	0
Ly_2	VVS	4	2	1	0	1
		0,211	0,167	0,1	0	0,333
	VS	0	1	2	0	0
		0	0,083	0,2	0	0
	S	0	2	1	0	0
		0	0,167	0,1	0	0
	M	3	2	1	2	2
		0,158	0,167	0,1	0,2	0,667
	L	4	3	1	4	0
		0,211	0,25	0,1	0,4	0
	VL	6	1	2	3	0
		0,316	0,083	0,2	0,3	0
	VVL	2	1	2	1	0
		0,105	0,083	0,2	0,1	0
Ly_3	VVS	0	0	0	0	0
		0	0	0	0	0
	VS	0	1	1	0	0
		0	0,083	0,1	0	0
	S	6	2	5	6	0
		0,316	0,167	0,5	0,6	0
	M	7	3	1	2	1
		0,368	0,25	0,1	0,2	0,333
	L	3	1	1	0	0
		0,158	0,083	0,1	0	0
	VL	0	0	0	0	0
		0	0	0	0	0
	VVL	3	5	2	2	2
		0,158	0,417	0,2	0,2	0,667

Table 9 Multi-valued logical-probabilistic function of the flotation process for double delay between inputs and outputs.

Models of complex processes under uncertainty using MLPF

3.2. Network Structure Model [8]

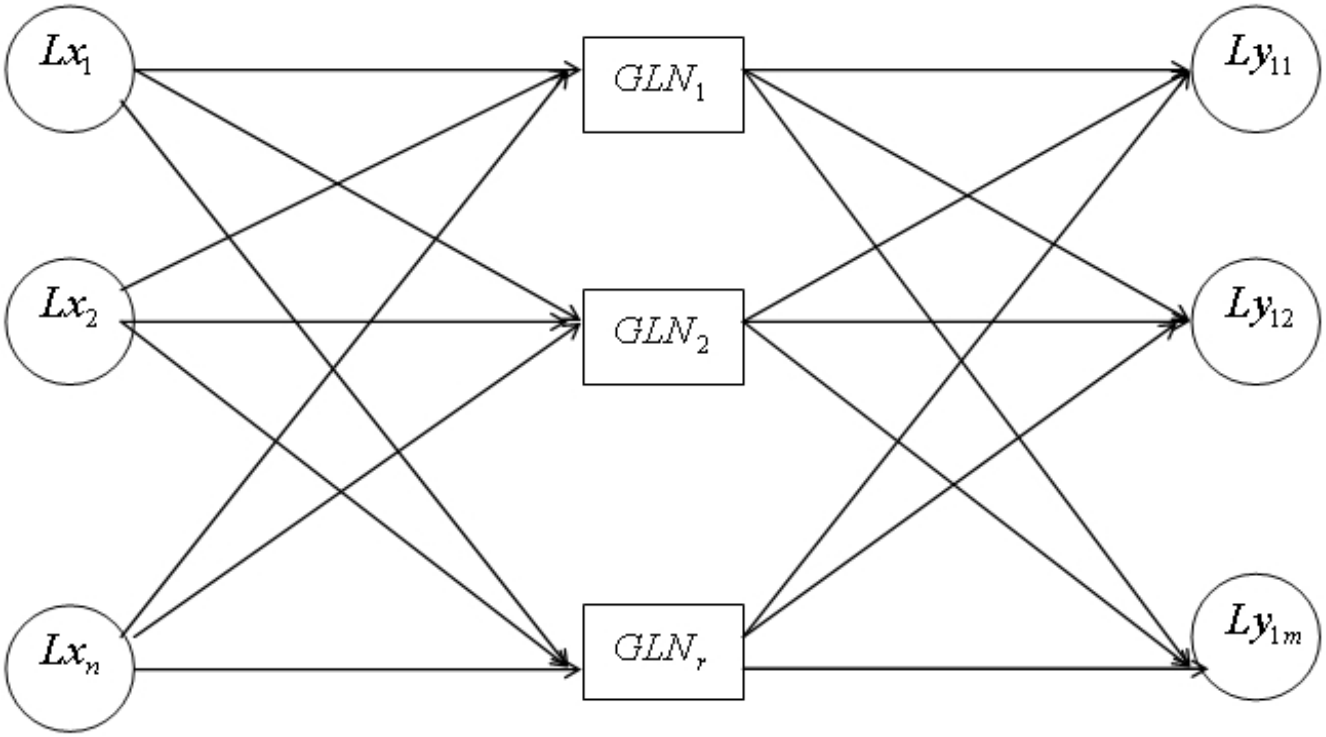

Figure 6 Model with network structure

Models of complex processes under uncertainty using MLPF

The following relations are introduced here [8]:

R_{LXGLNr} - is the relation between sets of the inputs and dominant grouping sequence sets ;

$R^*_{LX\ GLNr}$- is the relation between the frequency of occurrence of elements of the inputs and intermediate layers;

$R_{GLNrLyeq}$ – is the relation between the logical values of the elements of intermediate layer and output layer

$R^*_{GLNr\ Lyeq}$ - is the relation between the frequency of occurrence of the elements between the intermediate and the output layer.

Models of complex processes under uncertainty using MLPF

Using the network model, logical values and probability of occurrences of the outputs are calculated for each combination of logical values of measurable inputs [8].

$$Ly = R_{GLNrLy} \; X \; R_{LXGLNr} \; X \; LX \qquad (4)$$

$$p\{Ly\} = R^{*}_{GLNrLy} \; X \; R^{*}_{LXGLNr} \; X \; p\{LX\} \qquad (5)$$

Models of complex processes under uncertainty using MLPF

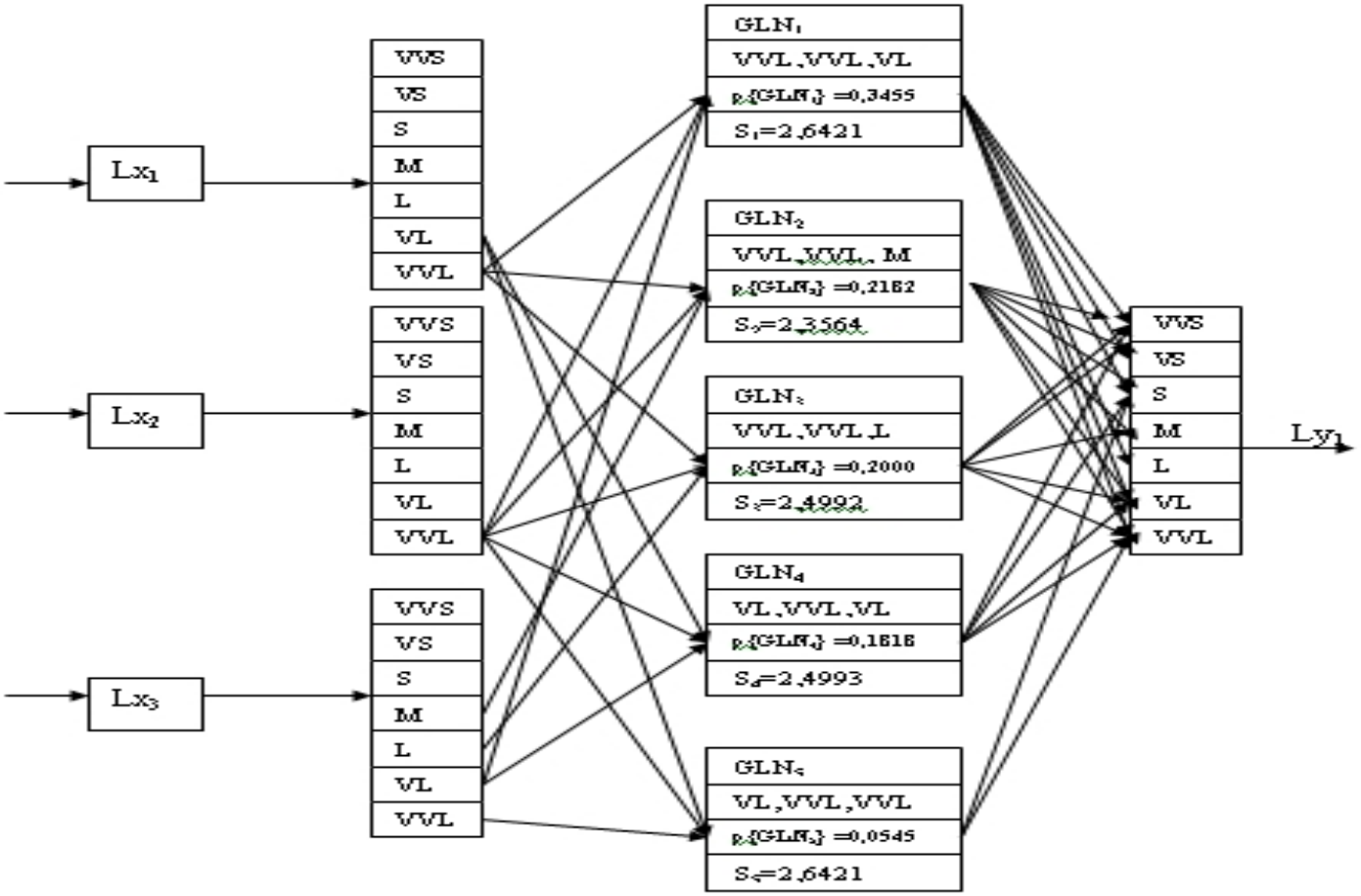

Fig.7 Network model $Ly_1=f(Lx_1,Lx_2, Lx_3)$ for 7 degree logic

Models of complex processes under uncertainty using MLPF

Using the network model there are calculated logical values for the output Ly_1 for each sequence data set for the inputs Lx_1,Lx_2 and Lx_3.

For example: If there are occur logical values for the input: Lx_1=VVL,Lx_2=VVL, Lx_3=M then it is activated grouping sequence set GLN_2, which activates the logical values of the output Ly_1 as follows, see Table 8':

VVS with frequency of occurrence p{VVS}=0,083,

VS with frequency of occurrence p{VS}=0,176

S with frequency of occurrence p{S}=0,333

M with frequency of occurrence p{M}=0,176

VL with frequency of occurrence p {VL}=0,176

VVL with frequency of occurrence p {VVL}=0,083

Models of complex processes under uncertainty using MLPF

Table 2 Relative values and their corresponding logic values

	VVS	VS	S	M	L	VL	VVL
min	0,0000	0,1429	0,2857	0,4285	0,5713	0,7141	0,8570
max	0,1428	0,2856	0,4284	0,5712	0,7140	0,8569	0,9997
mean value	0,0714	0,2142	0,3570	0,4998	0,6426	0,7855	0,9283

Models of complex processes under uncertainty using MLPF

Table 3 Relations between mean values for Lx_1 and dominant grouping sequence sets

GLNr	GLN_1	GLN_2	GLN_3	GLN_4	GLN_5
Lx_1	2,6421	2,3564	2,4992	2,4993	2,6421
0,9283	2,8461				
0,9283		2,5384			
0,9283			2,6922		
0,7855				3,1817	
0,7855					3,3635

Models of complex processes under uncertainty using MLPF

Table 4 Relations between mean relative values for Lx_2 and dominant grouping sequence sets

GLNr	GLN_1	GLN_2	GLN_3	GLN_4	GLN_5
Lx_2	2,6421	2,3564	2,4992	2,4993	2,6421
0,9283	2,8461				
0,9283		2,5384			
0,9283			2,6922		
0,9283				2,6923	
0,9283					2,8461

Models of complex processes under uncertainty using MLPF

Table 5 Relations between mean relative values for Lx_3 and dominant grouping sequence sets

GLNr	GLN_1	GLN_2	GLN_3	GLN_4	GLN_5
Lx_3	2,6421	2,3564	2,4992	2,4993	2,6421
0,7855	3,3635				
0,4998		4,7146			
0,6426			3,8892		
0,7855				3,1817	
0,9283					2,8461

Models of complex processes under uncertainty using MLPF

Table 6 Relations of frequency of occurrences between dominant grouping sequence sets GLNr and the output logic values Ly_1

$p(GLN_r)$	GLNr	VVS	VS	S	M	L	VL	VVL
0,3455	GLN_1	0,053	0,158	0,105	0,000	0,105	0,474	0,105
0,2182	GLN_2	0,083	0,167	0,333	0,167	0,000	0,167	0,083
0,2000	GLN_3	0,364	0,091	0,000	0,182	0,000	0,091	0,273
0,1818	GLN_4	0,200	0,000	0,200	0,000	0,000	0,500	0,100
0,0545	GLN_5	0,000	0,000	0,667	0,000	0,000	0,333	0,000

Conclusion

- Two MLPF based models for knowledge extraction from multi-factor, non-stationary, non-linear complex processes are proposed [7,8].
- The model with updatable knowledge base is illustrated with real data sets for an industrial process from the mining industry [7].
- The difference between the two models is that the model with updatable knowledge base uses knowledge extraction in the form of production rule whereas the model with network structure uses a network whose elements can perform computational logical operations [8].
- The model with network structure is better for non-stationary processes than the model with updatable knowledge base because of its capability to interpolate new data.

Acknowledgment

This research is supported by project

AComIn - "Advanced Computing for Innovation", grant 316087,
funding by FP7 Capacity Programme,
Research Potential of Convergence Regions (2012-2016)

Reference

1. Gray J., M. Research, J. Han, M. Kamber, “Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)”, Second Edition, Series Editors by Elsevier Inc.,2006.
2. Ruan D., G. Chen, E.Kerre, G.West (Edts.), „Intelligent Data Mining: Techniques and Applications (Studies in Computational Intelligence)”, Springer-Verlag, Berlin,Hidelberg, 2010.
3. Hopgood A., “Knowledge–Based Systems for Engineers and Scientists”, Second Edition, CRC-Press, pp.57-85, 1993.
4. Guida G., C.Tasso. “Design and Development of Knowledge Based Systems. From Life Cycle to Methodology”. John Wiley &Sons, pp.3-30, 1994.
5. Lee J. (Ed.), “Software Engineering with Computational Intelligence, Studies in Fuzziness and Soft Computing”, Springer, 2003.
6. Vatchova B. , “Derivation and Assessment of Reliability of Knowledge for Multifactor Industrial Processes”, PhD Thesis included 167 pages, Bulgarian Academy of Sciences, Sofia, 2009, (in Bulgarian).
7. GegovE.A.,B.Vatchova,E.D.Gegov,”Multi-valued Method for Knowledge Extraction and Updating in Real Time”, IEEE’04,Varna,Bulgaria,vol.2, pp. 17-6- 17-8,2008.
8. Vatchova B., A. Gegov. Knowledge Extraction Methods for Complex Processes Operating Under Uncertainty. IEEE 6th International Conference ‘Intelligent Systems’ 2012,978-1-4673-2278-2/12,volume II, pp.009-013 Sofia, Bulgaria.

Reference

9. Vatchova B. Logical Method for Knowledge Discovery based on Real Data Sets. IADIS conference on Data Mining, Rome, Italy, 24-26 July,2011 pp. 203-207, ISBN: 978-972-8939-53-3 © 2011 IADIS.
10. Gegov E., B.Vatchova, “Extraction of knowledge for complex objects from experimental data using functions of multi-valued logic”, European Conference on Complex Systems ‘09, University of Warwick, Coventry, UK, Sept.21-25, 2009.
11. Gegov E. , ”Methods and Applications into Computer Intelligence and Information Technologies of Control Systems, Publisher “St. Ivan Rilsky”, Sofia, 2003, (in Bulgarian).
12. Larose D., ”Data Minig Methods and Modles”, A .John Wiley&Sons, Inc.Publication, New Jersey, Canada,2006.
13. Han J., M.Kamber, “Data Mining Techniques”, Morgan Kaufmann Publisher, 2005.
14. Tim J.M., “Artifitial Intelligence A systems Approach” Infinity Science Press LLC, Hingham Massachusets, New Delhi,2007.
15. Kandel A., M.Last., H.Bunke, “Data Mining and Computational Intelligence“, Physical-Verlag, Heidelberg, 2001.