determining of attribute informativeness) and data transformation (normalisation, construction of new attributes and attribute aggregation), which are described in detail in [5]–[12]. Data pre-processing is one of the most resource-consuming processes – it can take up to 80 % time and work of all data analysis process [4].

In the analysis process of the present research, data are normalized using *z-score normalisation with standard deviation* because this method does not require defining minimum and maximum limits for attributes and it is considered to be one of the most popular data normalisation methods in data mining [13]. A normalized value of attribute A_i' can be calculated as follows (1):

$$A_i' = \frac{A_i - \bar{A}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - \bar{A})^2}}, \quad (1)$$

where

$\bar{A}$ – mean arithmetic value of attribute A;

A_i – i-th value of attribute A;

n – numbers of records in the data set.

Construction of new attributes is used in order to improve the process of data mining. Attribute construction is carried out by value merging using mathematical operations, e.g., attribute *Weight* and *Height* can be used to construct an attribute named *BMI* [14], which represents body mass index and is calculated as follows in (2):

$$BMI = \frac{Weight}{Height^2}, \quad (2)$$

where

Weight – weight of the respondent, kg;

Height – height of the respondent, m.

B. Attribute Selection

Attribute selection decreases the size of a data set and, therefore, increases the speed of data processing algorithm and accuracy of its results [5]. Attribute selection means determining a subset of the most informative attributes. Attribute selection consists of four steps: subset selection, subset evaluation, testing of end criterion and analysis of the obtained results that are shown in Fig. 1.

Selection of the most informative attributes is carried out using *CfsSubsetEval* method, which is considered one of the most popular attribute informativeness evaluation methods used in data mining [11]. Method *Cfs* (Correlation-Based Feature Selection) is an attribute evaluation method based on a correlation analysis.

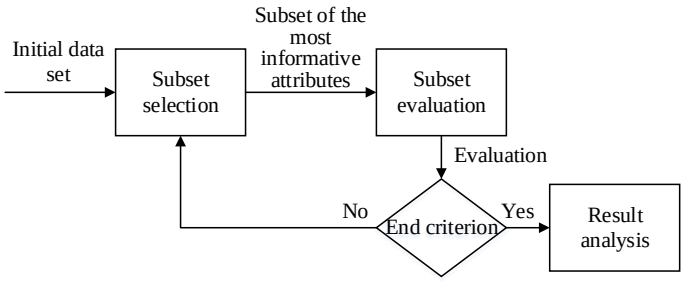

Fig. 1. Selection process of the most informative attributes.

The main idea behind it is to acquire a data set with attributes that have high correlation with class attribute and the smallest possible inter-attribute correlation. The obtained attribute subset is evaluated using (3), which is based on an Equation for Pearson correlation with standardized values [15]:

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}, \quad (3)$$

where

r_{zc} – correlation coefficient (evaluation metric);

k – number of attributes;

$\overline{r_{zi}}$ – mean correlation of attribute-class pairs;

$\overline{r_{ii}}$ – mean correlation of attribute-attribute pairs.

The search algorithm used with *CfsSubsetEval* attribute evaluation method is a *genetic algorithm*. *Genetic algorithm* is a search and optimisation approach, which is based on mechanisms of natural evolution like crossover, mutation and survival of the fittest (more optimal) individuals. Attribute selection using *genetic algorithm* is carried out similarly to a random search. It allows a population, formed by several individuals (solutions), to evolve based on a specific set of rules until a satisfactory result is achieved (being close to optimal result, which can be achieved in affordable time using the available resources). The *genetic algorithm* consists of three steps:

- selection – a set of best individuals is selected for a following generation;
- crossover – forms new individuals by combining pieces of good individuals that already exist in the population;
- mutation – changes individuals by inflicting random changes in order to increase diversity in the population.

Life-cycle of each generation ends by checking the existing population against the end criterion: if the criterion is not met, another population is evolved and changed using the same steps (described previously) and then evaluated. The cycle of the *genetic algorithm* repeats until a pre-defined end criterion is reached [16]. The working principle of the algorithm is depicted in Fig. 2.

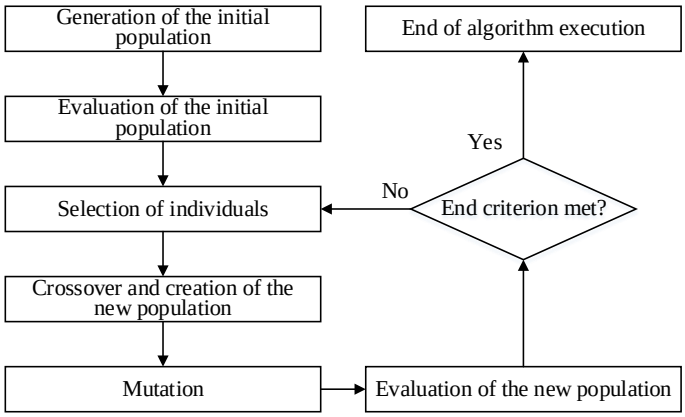

Fig. 2. Steps of the genetic algorithm.

In the beginning, the initial population is generated. It represents a set of possible solutions, where each solution is coded into a chromosome consisting of atomic genes that can be used to form any solution in the possible solution space. The next step involves the evaluation of the population and the use of the best individuals for next generation (the process of selection). Generation is a step in which the same population of individuals is used for operations in order to form new individuals that would be closer to the optimum. The next sub-steps include crossover and mutation using the individuals of the population, which result in new individuals with different genetic information (different solutions). When each new population is evaluated, it is tested against the end criterion. If the end criterion is met, the algorithm ends its operation. If the end criterion is not met, the algorithm continues with the selection of a new population for next generation [17].

Accuracy of the attribute informativeness evaluation is estimated using *k-fold cross-validation* [11].

C. Data Classification

Data classification is a process, which produces a model (classifier), which can be used to determine membership of observations to a group or class, by using only a data set with records with determined membership to groups or classes. Classification process consists of two steps:

- Training, where a training data set is used to train a classifier (or build it);
- Testing, where a test data set is used to test the accuracy of the classifier obtained in the first step. If the accuracy of the classifier is sufficient (as determined by system developers or experts), it can be used for classification of new records.

Evaluation of classifier accuracy is carried out using several metrics like *classifier overall accuracy*, *sensitivity* and *specificity*. Values of these metrics determine the suitability of each classifier for solution of the defined task [5], [17].

Simultaneously with the accuracy metric calculations, some approaches for result quality assurance have to be used as well, such as *k-fold cross-validation*. *K-fold cross-validation* divides a data set into *k* subsets by assigning a record to a subset randomly and keeping sizes of subsets equal. Classifier is trained and tested *k* times. Each time one subset is used for testing and the other *k*−1 subsets are used for training of a classifier. After each iteration, the results are entered into a confusion matrix, adding them to the results of previous iterations. The final result (confusion matrix and metrics calculated from it) includes results of all k iterations and estimates results of an algorithm as if it were tested on the whole available data set [11].

Experiments with classifiers include applications of the most popular classical classification algorithms [18]:

- *C4.5*, which uses inductive decision trees as classifiers and is based on algorithm *ID3* with extension that allows it to process data sets that include discrete, continuous and missing data [19];
- *CN2*, which uses inductive reasoning and in each iteration searches for a condition that would cover a large subset of objects with the same class value and as few objects from a different class as possible. When a condition is found, the algorithm removes objects, which are covered by the condition, from the data set and defines a rule that corresponds to the condition [20];
- *kNN* (*k-nearest neighbours algorithm*), which uses distance metric to determine closeness (distance) between each test object and each training set object and forms a *x*y* distance matrix for *x* objects of the training set and *y* objects of the test set. For each object from the test set, a set of *k* neighbours is found by selecting *k* closest training set objects (based on the smallest distance). The class for each test set record is determined based on a majority vote principle of the neighbour set with or without additional use of heuristics or weighted voting [21].

IV. Experiments and Result Analysis

Experiments were carried out using data from the database of a study carried out by the Latvian 'Interdisciplinary Research Group for Early Cancer Detection and Cancer Prevention'. This database was used to extract a data set with 136 attributes and 859 records. The class attribute for these experiments was the result of *H. pylori* test (positive or negative), which was carried out for each participant in the study. The selected attributes describe features of the individuals like nationality, education level, family situation, income level, occupation, smoking and alcohol consumption habits, physical activity, lifestyle, working environment, eating habits, family members, medical history, as well as age, gender, weight and height.

The experiments were carried out using *OrangeCanvas* data analysis tool with a unified experiment process model. This model provided attribute informativeness evaluation based on *CfsSubsetEval* method, using *genetic algorithm* based search approach. Classification was carried out using *C4.5*, *CN2* and *kNN* algorithms, evaluating the classifiers by *overall accuracy*, *sensitivity* and *specificity*. Classification results were evaluated using *10-fold cross-validation*.

The initial data set with 136 attributes obtained from the database was processed in order to increase its quality by

transforming, aggregating, merging and selecting the most informative results, leaving a data set with 65 attributes.

The experiments were carried out according to the process whose diagram is displayed in Fig. 3.

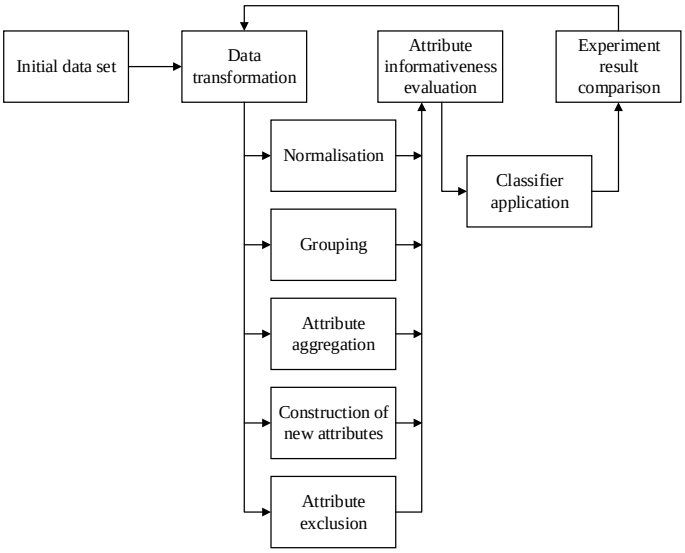

Fig. 3. Process of experiment execution.

The data set transformation involved several data pre-processing methods. Then the data set obtained in the pre-processing stage was processed by evaluating attribute informativeness and removing the least informative attributes. Then the data set with the determined most informative attributes was used in classification applying several classification algorithms. The obtained results were logged and mutually compared based on overall accuracy, sensitivity and specificity of the classifiers. These metrics were calculated from confusion matrices representing results of the various classifiers.

One of the attribute transformation processes is presented in Table I, where attributes *Height* and *Weight* were used to calculate and construct a new attribute *BMI*, which was used to aggregate a nominal attribute *BMI Categories*.

TABLE I

DATA TRANSFORMATION RESULTS

Height	Weight	BMI	BMI Categories
175	96	31	Obesity
175	108	35	Obesity
174	66	22	Normal weight
168	91	32	Obesity
185	103	30	Obesity
187	83	24	Normal weight
190	86	24	Normal weight
162	77	29	Overweight

Table II shows the process of several attribute grouping to derive a new attribute characterising *consumption intensity of smoked food*: *days per week* and *months per year*.

Table III shows experiment results acquired using different methods of data transformations. After the 3rd experiment, the number of data set records was decreased by removing respondents who had used medicine to eradicate infection caused by *H. pylori*.

TABLE II

DATA GROUPING RESULTS

Days per week	Months per year	Consumption of smoked food	Comment
0	0	0	None
1–2	1–4	1	Low
3–5	4–7	2	Medium
6–7	8–12	3	High

Further experiments were also carried out without the attributes that described the use of medicine for *H. pylori* eradication.

The best result of all experiments is the one achieved in the 16th experiment, which used *C4.5* classification algorithm – overall classifier accuracy of 68 %, sensitivity of 85 % and 29 % of specificity. This result was obtained in a data set with the following 20 attributes: *Age*, *Gender*, *Height*, *Weight*, *Nationality*, *Heavy metals*, *Volatile chemical substances*, *Organic chemical substances*, *Chemicals used in agriculture*, *Vibration*, *Radiation*, *Biological factors*, *Smoking*, *Alcohol consumption*, *Walking*, *Moderate activity*, *Workload*, *Salty food*, *Smoked food* and *Diseases*. Analysis of these results shows that development of *H. pylori* infection is promoted by hazardous habits described in literature previously: smoking, alcohol consumption, medical history and increased consumption of salt. Other factors determined in this research are the following: work environment and level of physical activities. The experiments show that hazardous work environment largely contributes to developing *H. pylori* infection. A significant part of attributes (38 %) left in the final set are connected to the working environment. The final data set also holds the attribute *Workload*, which means that it also contributes to the risk of developing *H. pylori* infection, mainly due to the caused stress and lifestyle changes.

In order to make the results more demonstrative, relationships among attributes and respondents with positive H. pylori tests were calculated. Part of this information is presented in Fig. 4. It shows that 42 % of the respondents with positive tests are in the age range of 40 to 50 years, 25 % of the respondents regularly face volatile chemical compounds at work and 27 % face vibration at work. 41 % of these respondents smoke, 38 % of these respondents do not carry out any physical activities, 28 % have high workload.

Most of the respondents with positive test results (75 %) consume alcohol and 23 % of these respondents have had gastric diseases. These statistics leads to a conclusion that one of the highest impacts on developing *H. pylori* infection is from hazardous factors in the work environment, such as volatile chemical compounds and vibration. A large number of respondents use alcohol, which is a known and previously poven factor in *H. pylori* infection development. A little less than half of the respondents are smokers (41 %) and physically passive(39 %).

TABLE III
EXPERIMENT RESULTS

Data set	No. of attributes	No. of records	*C4.5*			*kNN*			*CN2*		
			Acc.	**Sen.**	**Spec.**	**Acc.**	**Sen.**	**Spec.**	**Acc.**	**Sen.**	**Spec.**
Initial data set	65	859	0.62	0.79	0.28	0.54	0.61	0.42	0.64	0.90	0.16
1. Experiment	64	859	0.60	0.75	0.30	0.57	0.71	0.30	0.65	0.20	0.90
2. Experiment	56	859	0.60	0.75	0.30	0.55	0.69	0.28	0.63	0.87	0.24
3. Experiment	56	859	0.61	0.71	0.40	0.55	0.67	0.28	0.64	0.86	0.19
4. Experiment	50	742	0.62	0.72	0.42	0.55	0.68	0.30	0.63	0.86	0.16
5. Experiment	47	742	0.62	0.78	0.21	0.57	0.70	0.26	0.66	0.90	0.09
6. Experiment	27	742	0.61	0.79	0.17	0.57	0.67	0.31	0.65	0.88	0.10
7. Experiment	24	742	0.65	0.84	0.20	0.56	0.68	0.27	0.66	0.91	0.05
8. Experiment	9	742	0.70	0.96	0.07	0.65	0.78	0.32	0.71	0.98	0.04
9. Experiment	43	742	0.64	0.80	0.32	0.60	0.73	0.33	0.67	0.95	0.11
10. Experiment	40	742	0.64	0.82	0.22	0.63	0.77	0.27	0.70	0.95	0.07
11. Experiment	35	742	0.68	0.88	0.21	0.59	0.74	0.32	0.63	0.83	0.13
12. Experiment	30	742	0.66	0.81	0.27	0.62	0.77	0.25	0.65	0.87	0.12
13. Experiment	28	742	0.65	0.81	0.28	0.62	0.62	0.22	0.64	0.86	0.13
14. Experiment	26	742	0.67	0.85	0.25	0.65	0.78	0.33	0.64	0.87	0.06
15. Experiment	23	742	0.67	0.82	0.31	0.60	0.75	0.24	0.66	0.90	0.10
16. Experiment	20	742	**0.68**	**0.85**	**0.29**	0.63	0.77	0.27	0.65	0.90	0.04

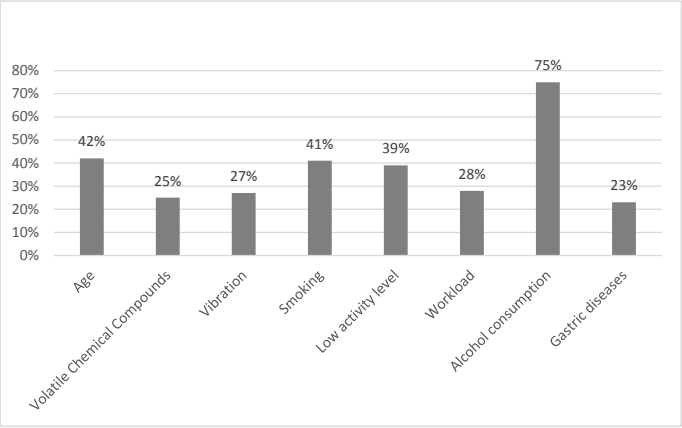

Fig. 4. Analysis of the obtained results.

V. CONCLUSION

Experimental evaluation of the initial data set led to decreasing of the data set to 20 attributes and 742 records. Most of the experiments were carried out using data pre-processing methods and constructing different variations of the data set to determine the most suitable attributes or attribute subsets.

The experimental analysis of factor sets and data set sizes showed that these changes had mostly small impact on overall accuracy of the classifiers. The overall accuracy for *C4.5* varied in the range from 60 % to 70 %. Most of the variations were due to strong variations in sensitivity and specificity. Since the goal of the research was to determine factors that influence positive class, sensitivity was given the major importance but the desired specificity should be around 30 % or one correct prediction among three to avoid assigning all individuals to the positive class and receiving 0 % specificity and no useful information from the model.

The experimental evaluation determined the most suitable classifier, which was constructed using *C4.5* classification algorithm. This classifier showed that an infection risk was influenced by factors contributing to *H. pylori* development like *working environment* and *physical activity*. Closer analysis of these factors showed that respondents, who had positive *H. pylori* test results, worked in an environment where they faced either *vibration* or *volatile chemical compounds*, as well as heavy *workload*. Another contributing factor was *lack of physical activity*, which was confirmed in 38 % positive respondents who did not carry out any physical exercise.

In cases when a person faces hazardous work, environment and volatile chemical compounds at work, a doctor should appoint them to *H. pylori* test, which could facilitate early detection and prevention of gastric diseases and cancer.

ACKNOWLEDGMENT

The present research has been developed by the GISTAR within the project “Interdisciplinary Research Group for Early Cancer Detection and Cancer Prevention” and supported by the International Agency for Research on Cancer (IARC).

REFERENCES

[1] *Helicobacter Pylori and Cancer*. USA: National Cancer Institute, 2015. [Online]. Available: http://www.cancer.gov/cancertopics/factsheet/Risk/h-pylori-cancer. [Accessed September 10, 2015].

[2] W. D. Chey, B. C. Wong, "American College of Gastroenterology guideline on the management of Helicobacter pylori," *American Journal of Gastroenterology*, vol. 102, pp. 1808-1825, 2007. http://dx.doi.org/10.1111/j.1572-0241.2007.01393.x

[3] L. E. Wroblewski, R. M. Peek, K. T. Wilson, "Helicobacter pylori and Gastric Cancer: Factors That Modulate Disease Risk," *Clinical Microbiology Reviews*, vol. 4, pp. 713–739, 2010. http://dx.doi.org/10.1128/CMR.00011-10

[4] *Study to Prevent Gastric Cancer Mortality*. Latvia: GISTAR, 2015. [Online]. Available: https://www/gistar.eu. [Accessed September 11, 2015].

[5] Y. Zhu, X. Zhou, J. Wu, J. Su, G. Zhang, "Risk Factors and Prevalence of Helicobacter pylori Infection in Persistent High Incidence Area of Gastric Carcinoma in Yangzhong City," *Gastroenterol Res Pract.*, 2014:481365, 2014.

[6] M. P. Dore, H. M. Malaty, D. Y. Graham, G. Fanciulli, G. Delitala, G. Realdi, "Risk Factors Associated with Helicobacter pylori Infection among Children in a Defined Geographic Area," *Clin Infect Dis*, Vol. 35 (3), pp. 240-245, 2012. http://dx.doi.org/10.1086/341415

[7] The EUROGAST Study Group. Epidemiology of, and risk factors for, Helicobacter pylori infection among 3194 asymptomatic subjects in 17 populations. The EUROGAST Study Group. *Gut*, vol. 34(12), pp. 1672-1676, 1993. http://dx.doi.org/10.1136/gut.34.12.1672

[8] S. H. Lim, J. W. Kwon, N. Kim, G. H. Kim, J. M. Kang, M. J. Park, J. Y. Yim, H. U. Kim, G. H. Baik, G. S. Seo, J. E. Shin, Y. E. Joo, J. S. Kim, H. C. Jung, "Prevalence and risk factors of Helicobacter pylori infection in Korea: Nationwide multicenter study over 13 years," *BMC Gastroenterology*, 13:104, 2013. http://dx.doi.org/10.1186/1471-230X-13-104

[9] M. Hasosah, M. Satti, A. Shehzad, A. Alsahafi, G. Sukkar, A. Alzaben, A. Sunaid, A. Ahmed, S. AlThubiti, A. Mufti, K. Jacobson, "Prevalence and Risk Factors of Helicobacter pylori Infection in Saudi Children: A Three-Year Prospective Controlled Study," *Helicobacter*, vol. 20(1), pp. 56–63, 2015. http://dx.doi.org/10.1111/hel.12172

[10] J. Han and M. Kamber, *Data Mining: Concepts and Techniques. Second Edition*. Morgan Kaufmann, Elsevier Inc., 2006.

[11] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques – 2nd edition.* Amsterdam etc.: Morgan Kaufman, 2005.

[12] D. Pyle, *Data Preparation for Data Mining.* San Francisco etc.: Morgan Kaufmann, 1999.

[13] P. N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*. Boston: Pearson Addison-Wesley, 2006.

[14] *Calculate Your Body Mass Index*. USA: National Heart, Lung and Blood Institute, 2015. [Online]. Available: https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm. [Accessed September 11, 2015].

[15] M. A. Hall, "Correlation-based feature selection for machine learning," Doctoral Thesis, Hamilton: University of Waikato. 1999.

[16] R. Tiwari, M. P. Singh, "Correlation-based Attribute Selection using Genetic Algorithm," *International Journal of Computer Applications*, vol. 8, pp. 28-34. 2010. http://dx.doi.org/10.5120/847-1182

[17] A. Bharathi, E. Deepankumar, "Survey on Classification Techniques in Data Mining," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 7, pp. 1983–1986. 2014.

[18] X. Wu, V. Kumar, J. R. Quinlan, et al., "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, pp. 1-37. 2007. http://dx.doi.org/10.1007/s10115-007-0114-2

[19] J. R. Quinlan C4.5: Programs for Machine Learning. *San Mateo: Morgan Kaufmann Pub.*, 1993.

[20] P. Clark, T. Niblett, "The CN2 induction algorithm," *Machine Learning*, 3(4), pp. 261-283, 1989. http://dx.doi.org/10.1007/BF00116835

[21] M. W. Berry, M. Browne, *Lecture Notes in Data Mining.* World Scientific Publishing Co. Pte. Ltd., 2006.

Arnis Kirshners, *M. sc. ing.*, is a Doctoral Student, Lecturer, Department of Modelling and Simulation, Faculty of Computer Science and Information Technology, Riga Technical University, and a Researcher, Faculty of Medicine, University of Latvia. He received his diploma from the Department of Modelling and Simulation from Riga Technical University. His research interests include data mining and knowledge extraction, intelligent systems, programming and database. He has 18 publications in the area.
Contact information: 1 Kalku Str., Riga, LV-1658, phone: +371 67089530.
E-mail: arnis.kirsners@rtu.lv

Inese Polaka, *Dr. sc. ing*,. is a Lecturer at the Institute of Information Technology of Riga Technical University (Latvia) and a Leading Researcher at the Faculty of Medicine, University of Latvia. Main research interests include data mining, machine learning, classifiers, evolutionary algorithms and their applications, as well as bioinformatics and biostatistics.
Contact information: 1 Kalku Str., Riga, LV-1658, phone: +371 67089530.
E-mail: inese.polaka@rtu.lv

Ludmila Aleksejeva received her *Dr. sc. ing.* degree from Riga Technical University in 1998. She is an Associate Professor at the Department of Modelling and Simulation, Riga Technical University. Her research interests include decision making techniques and decision support system design principles, as well as data mining methods and tasks, and especially collaboration and cooperation of the mentioned techniques. She has more than 30 academic and scientific publications.
Contact information: 1 Kalku Str., Riga, LV-1658, phone: +371 67089530.
E-mail: ludmila.aleksejeva_1@rtu.lv