Data Mining Techniques in Simulation Results Analysis

Irina Sitova
Department of Modelling and Simulation
Riga Technical University
Riga, Latvia
irina.sitova@rtu.lv

Jelena Pecerska
Department of Modelling and Simulation
Riga Technical University
Riga, Latvia
jelena.pecerska@rtu.lv

***Abstract*— The research is carried out in the area of analysis of simulation results by using data mining techniques. The goal of this research is to explore the applicability of data mining techniques in the area of simulation result analysis, offer data mining techniques application scheme in analysis of simulation results, as well as demonstrate the usage of these techniques in analysis of experimental data. As a result of the theoretical study, an approach is proposed, consisting of two stages and combining the fundamental principles of data farming and knowledge discovery. A variety of data mining techniques, including correlation analysis, clustering and several visualization mechanisms of results are used for the knowledge discovery. The approach is applied to the analysis of experimental data. The performence of a queueing system was analyzed, and knowledge and decision rules are obtained from simulation results.**

Keywords—Data mining, discrete-event system simulation, simulation results analysis, queueing system.

I. INTRODUCTION

The purpose of the most simulation projects is obtaining information about the behaviour of the system under consideration in different conditions. Researchers obtain this information after experiments with a verified, calibrated and validated model. In this paper, the discussion is limited to the results obtained from discrete-event system (DES) simulation models, which are most effectively used for the analysis of dynamics of complex artificial material systems.

The traditional simulation report is a generalization of the output statistics after running the simulation model. The output statistics is based on the results of a planned experiment or series of experiments. "Output analysis is the examination of data generated by simulation" [1]. The output statistics is the main "trophy" of the researcher-simulationist and allows understanding of the behaviour of the system, to formulate forecasts, to compare alternatives or to solve the optimization tasks of system parameters. The question under consideration is: How should one perform data analysis and results interpretation? The traditional approach to simulation output analysis implies replication design, estimation of performance indicators and analysis of system behaviour, based on these estimations. However, this final analysis cannot be reduced to statistical methods only. Various researchers are investigating simulation results combining statistical analysis and data mining [4], [5], [6], [7], thus providing more efficient and versatile output analysis, and dealing with potentially huge amount of simulation output data.

The authors provide a review of approaches in the area of analysis of simulation results by using data mining techniques. The goal of this research is to explore the applicability of data mining techniques in the area of simulation result analysis and to introduce data mining techniques application scheme for analysis of simulation results. As a result of the theoretical study, a two-stage approach is formulated, combining the fundamental principles of data farming and knowledge discovery. A variety of data mining techniques, including correlation analysis, clustering and several visualization mechanisms of results are used for the knowledge discovery. The developed approach is applied to the analysis of experimental data of a simple DES simulation model.

We hypothesized that data mining techniques may provide the better interpretation of simulation output as well as visualization of outputs. Another issue was to make sure that data mining reveals not only trivial knowledge from simulation output. Finally, knowledge and decision rules are obtained from simulation results coupled with the relevant visualization.

II. RELATED WORK

A. Heuristic-Based Searches with Simulation Models, 2001

One of the pioneers in field of integrating data mining techniques in simulation results analysis were Brady and Bowden [2]. Authors used heuristic-based searches with simulation models for solving one of the most important simulation challenges – selecting the optimal set of input variables and their values. There is the rule – a compromise between adding all important input variables and not overloading the model must be found. The provided solution is named an 'external' optimization, as soon as input variable determining process is made outside of the simulation model. Authors' method can provide 'better' answers than trial and error methods, when selection of the optimal set of input variables is made based on a large number of experiments. However, the authors mention the disadvantage of the proposed method, which is associated with the fact that people, who may not be accurate enough or even wrong, implement the most important phase – the selection of decision variables for the optimization. As a result, the method is excessively dependent on the decision makers that define the input variables.

B. A new form of computer simulation output, 2005

In contrast to [2] authors of [3] developed an internal approach for selecting optimization variables, that analyses the dynamics of input variables interaction within the simulation model. The aim of this research was to create a method that uses only one replication results of the simulation model to order the elements according to their importance. Simulation model elements including resources,

978-1-7281-0098-2/18/$31.00 ©2018 IEEE

entities, statuses, and their relationship displays the logic of model. During one replication the information that appears in the trace file is analysed. Each simulation model element has code or keyword for easier information interpretation from trace file. The frequency analyser program evaluates keywords appearance frequency in a trace file. Then, considering the information obtained, a correlation analysis was performed using the cosine method. The method described in [3] was applied to the practical problem, where the information obtained was considered non-trivial and it was transforms into useful knowledge that helped to improve the semiconductor manufacturing process. To conclude it is appropriate to use the proposed method when the simulation model consists big number of elements and it is necessary to know their interconnectedness.

C. *Combined Techniques Including Simulation, Data Mining and Knowledge Discovery, 2006*

The method described in [7] combines simulation, data mining and knowledge discovery techniques for optimizing aircraft engine (both short-term and long-term) life cycle costs (LCC). The method is based on the following algorithm – obtained data on operational functionality and costs from the simulation model are mined in order to discover the parametric relations that best describe the LCC value changes for each accepted strategic decision.

Authors in [7] use several data mining methods, such as linear regression, clustering and classification, to determine significant cost drivers. Two software tools are used: Arena as a simulation tool and the PDP as a data mining tool.

To test the effectiveness of the method, a case study was carried out. First, a linear regression model was constructed (with aim) to find the values that affect the LCC increase.

Then for the analysis of the parameters that affect the LLS value, the methods of classification were used. Classification was performed using decision trees-based algorithm - CART. Finally, the simulation results were analyzed using clustering algorithm – K-means algorithm. The main result of this work [7] was the proof that there is an ability to use data mining methods to identify and describe the cost drivers, obtaining them from the simulation results.

D. *Integrating data analytics and simulation methods, 2015*

Another research in the area [6] provides a methodology, according to which, variables that most effectively increase investigated system performance are extracted from data using data analytics methods. These parameters are used to prepare the appropriate simulation input data for execution of the scenarios. System optimization is performed based on simulation output data.

Two features distinguish the described methodology [6] from traditional simulation and optimization approaches a) the input data which is collected from various systems using smart devices (e.g., sensors) is large in scale, include different parameters and it is constantly replenished; b) usage of data mining methods, including association and classification methods, to identify more relevant variables that are related to specific performance indicators.

The authors of [6] demonstrate the application of the methodology for metal products manufacturing system model. As a result, a specific process plan is defined, which optimizes production costs and the methodology that integrates data mining, simulation and optimization helps to make more constructive decisions.

E. *Further development of data mining based approaches, actualities*

The authors of [4] are developing and testing the proposed approach by formulating applicable management strategy in [5]. The sequential application of data mining techniques – correlation analysis, clustering and best cluster selection – resulted in finding of prevailing input parameter values. The obtained information is used for decision about changing system parameter values. The previously formulated approach is successfully implemented for real world problem solution.

There is a considerable number of articles and projects concerning the merging of the concepts of data mining techniques in simulation results analysis. However, the scheme of application of data mining technology may differ from project to project and a sustainable approach in the area has not yet emerged. The following case study tends to contribute into obtaining knowledge from simulation results by applying the data mining technique.

III. Case Study. Queueing System Simulation and Results Analysis

A. *Queuing system with single server, a queue with infinite capacity, and several random factors*

The choice of the system for the case study is determined as follows: the system should be simple enough, yet the simulation results are not interpreted unambiguously – objective function alternatives of input variable sets are under consideration.

Thus, the system is a queuing system with queue with infinite capacity, exponential distribution of interarrival times, variable average value of this exponential distribution at different daytime, random service time, queue-length based probability of lost customers. Interarrival times and service times are independent. Service time is a function of demand for product, and demand distribution is described as empirical one. Probability of lost customers is described as a function of queue length. The detailed analytical study of system performance measures is complicated. For the purposes of performance analysis a discrete event simulation model of a system is created, providing the estimates of relevant performance measures: statistics of customers in system, time spend in system, occupation rate of the server, which part of the customer will be served, satisfied demand for product and some other performance measures.

The conceptual model of queueing system under consideration is shown in Fig. 1.

The goal of the case study is to find the combination of input variable values and adjustable timetables of work, providing the best values of the composite objective function.

B. *The research methodology*

To achieve the goal of the study, the process is consistently implemented as follows. The simulation model

of the system under consideration is constructed, simulation experiments are designed, replicated and results obtained. The simulation results are analyzed using data mining techniques. The domain of aacceptable values of experimental factors is revealed and visualized. The methodology is summarized in figure 2.

C. *Software implemented for the case study*

It can be stated that the expected performance indicators estimations are provided by any of the discrete-event simulation tools. Therefore, one of the widely used simulation packages was applied to create a simulation model. For the analysis of simulation results, the R software environment was used as a powerful data mining tool with available visualization.

D. *Simulation experiments and statistical results*

For obtaining data for mining 28 model experiment scenarios were developed with the realistic experimental factor values. Each of scenarios was replicated 25 times, providing 700 experiment results sets.

Figure 3 shows an overview of the model layout, a snapshot of some simulation results and the results table used for further analysis.

E. *Data mining technique application for simulation results analysis*

In frame of the research, the data mining techniques were applied to analyse corresponding data sets. The general scheme of Knowledge discovery is shown in figure 4. Knowledge discovery procedure implementation consists of 3 steps.

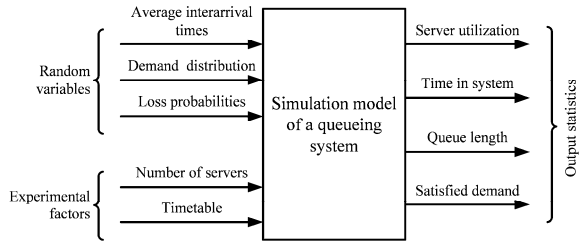

Fig. 1. Conceptual model for the case study

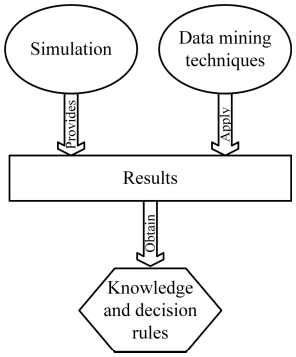

Fig. 2. Research methodology

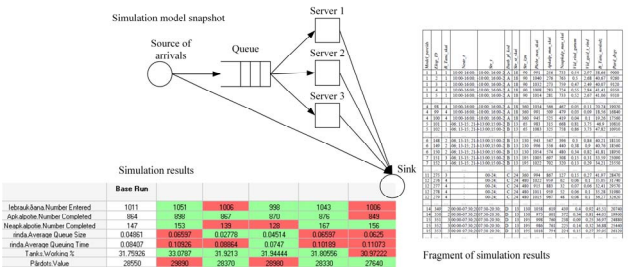

Fig. 3. The simulation model layout, a fragment of simulation report and a fragment of results summary used for data mining

F. *Results of corellation analysis*

Two output variables, with a strong negative correlation were defined as the target variables. As well as another important output variable that is correlated with the first two was considered. The summary of correlation analysis is provided in figure 5.

G. *Clustering*

The *k*-means clustering method was implemented for results partition into clusters. At the beginning the best number of clusters was determined, which is equal to 4, then the clustering was carried out, and finally the cluster ranking was performed. Clustering was carried out for three output variables, therefore the results can be vosualized in a three-dimensional graphics. *Plotly for R* library was used for this purpose. The visualization of obtained clusters is provided in figure 6.

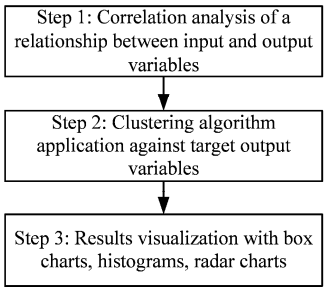

Fig. 4. Data mining techniques application scheme in analysis of simulation results

	B_Tank_skai	Str_st_skai	Str_izm	Piebr_mas_skai	Apkalp_mas_skai	Neapkalp_mas_skai	Vid_rind_garum	Vid_gaid_t_rind	B_Tank_noslodz	Pard_degv
B_Tank_skai	1.00									
Str_st_skai	0.00	1.00								
Str_izm	0.89	0.42	1.00							
Piebr_mas_skai	0.04	0.00	0.03	1.00						
Apkalp_mas_skai	0.83	0.17	0.81	0.09	1.00					
Neapkalp_mas_skai	-0.82	-0.17	-0.81	0.07	-0.99	1.00				
Vid_rind_garum	-0.91	0.03	-0.80	0.01	-0.76	0.76	1.00			
Vid_gaid_t_rind	-0.88	-0.04	-0.79	-0.02	-0.82	0.82	0.96	1.00		
B_Tank_noslodz	-0.81	0.23	-0.64	0.03	-0.41	0.41	0.86	0.73	1.00	
Pard_degv	0.83	0.17	0.81	0.09	1.00	-0.99	-0.76	-0.82	-0.40	1.00

Fig. 5. The summary of corellation analysis

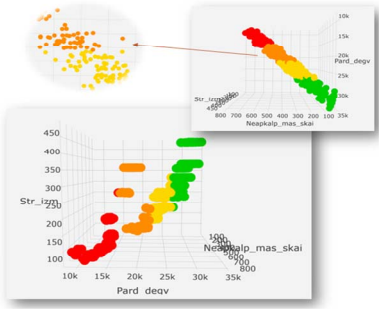

Fig. 6. The results of cluster analysis

H. Results visualization

The visualization goal of clustering results is to support the detection of input variable values that are specific for a particular cluster. To detect these values the box diagrams of two numeric input variables were created. The results obtained may be summarised in a table, giving the input variable ranges corresponding to each cluster. The box diagram of a single input variable is provided in figure 7.

The next stage of data visualization is carried out through input variable value distribution analysis in clusters. As a result the value hystograms were created for making conclusion about the most efficient working timetable. An example of input variable histogram is shown in figure 8.

The results of this stage may be summarized in a table, and provide a dominant values of this input variable in appropriate cluster. By using histogram, it is possible to exclude one of the timetables from further analysis as unefficient one.

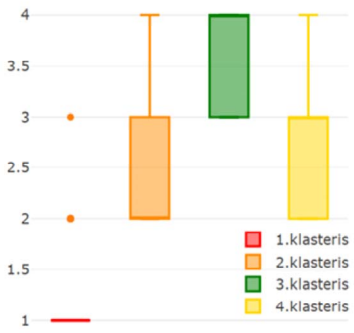

Fig. 7. The box diagram of a single input variable

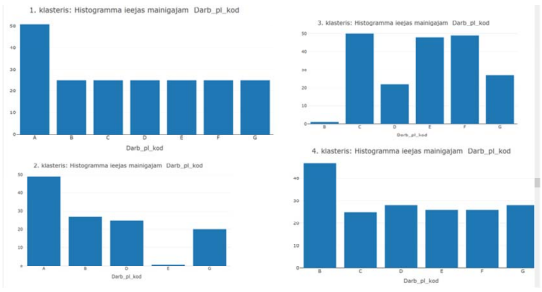

Fig. 8. The histograms of a single input variable values in clusters

Next, the radar diagrams were created for all the clusters showing average values of input and output variables. The radar diagrams for particular experiments. This visualization gives possibility to analyze several variable averages at a time. The radar diagram of one experiment is shown in figure 9.

Analyzing the radar diagram it is possible to make sure that one of the clusters is a target cluster.

Based on the experimental results the decision principles for the case study problem were formulated for detecting the input variable values that are specific for each cluster and in particular for the best cluster. Thus, a main goal of the case study – the best performance of a queueing system – was discovered.

CONCLUSION

A variety of data mining techniques, including correlation analysis, clustering and several visualization mechanisms of results, were applied during the knowledge discovery. Three experiments that correspond to defined combination of data mining techniques were designed. The results of the experiments supported the identification of the relationships between the simulation model input and output variables and definition of target output variables. All data records were arranged into groups according to the values of target output variables and the best group was chosen. According to experiment results, the values of the input variables that are typical for each created group, particularly for the best group, were defined. Due to this, knowledge and decision rules were obtained from simulation results.

Finally, it was concluded that:

1. The target output variable selection, which is carried out by clustering, is a decisive phase. This selectiondetermines the results of further analysis, as well as the speed and efficiency of the procedure in general.

2. If there are no strictly defined target values of the output variables, the most relevant technique may be clustering.

The simulation experiments performed in this research made it possible to obtain useful knowledge from simulation, and proved the initial hypothesis about the suitability of the proposed scheme for efficient simulation results analysis, knowledge discovery and decision formulation. The scheme is applicable for problem solving in queuing systems as well as in other simulation-based analysis projects.

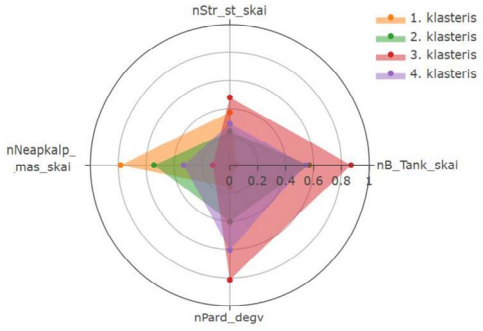

Fig. 9. The radar diagram of input and output variables of four clusters in a single experiment

REFERENCES

[1] Banks, J., Carson II, J. S., Nelson, B. L., & Nicol, D. M. (2010). Discrete-event System Simulation. Pearson.

[2] Brady, T. F., Bowden, R. A. The effectiveness of generic optimization routines in computer simulation languages. Proceedings of the 2001 Industrial Engineering Research Conference, 2001.

[3] Brady, T. F., Yelling, E. Simulation data mining: A new form of computer simulation output. Proceedings of the 2005 Winter Simulation Conference, 2005.

[4] Feldkamp, N., Bergmann, S., Strassburger, S. Knowledge Discovery in Manufacturing Simulations, Proceedings of the 3rd ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, 2015.

[5] Feldkamp, N., Bergmann, S., Strassburger, S. Knowledge discovery in simulation data: A case study of a gold mining facility, Proceedings of the 2016 Winter Simulation Conference, 2016.

[6] Kibira, D., Hatim, Q., Kumara, S., et al. Integrating data analytics and simulation methods to support manufacturing decision making, Proceedings of the 2015 Winter Simulation Conference, 2015.

[7] Painter, M.K., Erraguntla, M., Beachkofski, B., et al. Using simulation, data mining, and knowledge discovery techniques for optimized aircraft engine fleet management, Proceedings of the 2006 Winter Simulation Conference, 2006.