Peter BUBENÍK, Filip HORÁK

INTEGRATING SIMULATION INTO DATA MINING

1. Introduction

Data mining is a new technology of discovering knowledge in large volumes of data. [10] Many companies recently recognized this technology as one with a huge potential in respect to performance of industries in general, and individual companies in particular. Data mining is now being used in many fields of production for acquiring and deployment of knowledge in predictive maintenance, manufacturing, scheduling and various systems dedicated to decision support. Data mining approach can be used to extract and identify hidden patterns in large data sets and for discovering knowledge, which can be applied to solve individual problems in practice. [4] One of the most popular processing standard for data mining is CRISP-DM (CRoss Industry Standard Process for Data Mining), which will be described in more detail within following chapter.

Simulation approach is far more older than data mining, and nowadays, its usage in production companies is becoming still more and more likely, because it is one of the most popular methods for complex production systems analysis. [7] Using computer simulation has lead to many benefits in manufacturing practice, as are for example lowering the risk, increasing overall understanding of the system, lowering operational costs, shortening lead times, etc. [9]

Simulation, as well as data mining is a powerful tool. Their separate usage enables companies to take better and more relevant decisions. However, if combined, we can get even stronger tool for management based on verifiable knowledge. [8] From a data mining point of view, simulation can offer a sufficient volume of data needed for further processing, but above-mentioned standard CRISP-DM, does not refer to it in more detail. The aim of this article is to define a place for simulation in data mining process framework in the way that this technology could be used with better results and in bigger amount of individual cases that can arise in production practice.

2. Knowledge-Based Systems

Information systems that follow the procedural path and compute desired results based on fixed internal algorithms can be called traditional systems, as opposed to knowledge-based systems (KBS) which will be discussed more thoroughly over the course of this article. Main significant difference between traditional systems and now emerging knowledge-based systems is the knowledge-base, which is exploited by inference engine in order to find solution for the problem, which user of such system currently faces. These kind of problems would traditionally be solved by expert in target domain, but instead, knowledge-based systems are being developed, among the other things, in order to formalize, apply and preserve acquired knowledge so it stays in company. This is why knowledge-based systems are by many also called expert systems, and these terms can therefore be often used interchangeably.

The architecture of above mentioned KBS systems differs from traditional architecture in many aspects as seen in Figure 1. Basic idea of the system is that user communicates with inference engine through user interface, and specifies characteristics of his current situation. Inference engine acts as reasoning element, and tries to apply available knowledge from knowledge-base in order to provide user with a suitable solution to the problem.

Creating knowledge-based system is a complex domain-specific task and it is hard to define this process in terms of what precise steps must be followed. Nevertheless, it is possible to conclude, that creating user interface and the inference engine is a task that should require simple to advanced programming skills. This is due to the fact that user interface is nowadays an inseparable part of almost every information solution and an inference engine creation can be accelerated by already existing programming libraries such as Inference Engine Component Suite for Delphi or full free modifiable inference engines like Simple Rule Engine for .NET and many more.

In case of an actual knowledge-base, the knowledge acquisition (KA) process largely depends on selected domain in which the knowledge-based system would provide decision support.

In practice, there are two broader kinds of approaches. One can be described as an effort to formalize knowledge of company employees or experts in specified field. In another words, it is rather a process of converting tacit knowledge to explicit knowledge. The other approach is to derive this knowledge from historical data. This approach is called data mining.

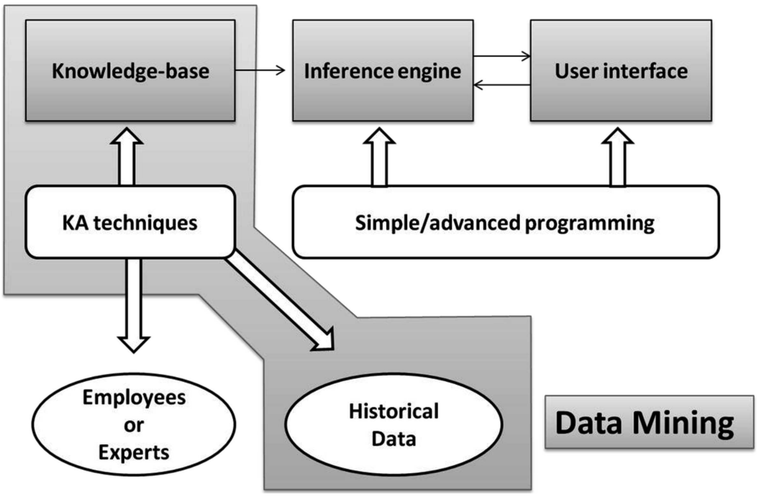

Fig. 1. KBS Architecture and its requirements

3. Data Mining

Data mining can be characterized as an interdisciplinary subfield of computer science. It is a computational process and its goal is usually to find patterns in large data sets. These patterns can be presented, stored and used as actual knowledge, and therefore aid decision processes. Disciplines that are considered to have a large impact on data mining are:

- artificial intelligence (AI)
- machine learning (usually also considered as part of AI)
- statistics
- database systems

3.1 Data mining in manufacturing environment

There have been numerous examples of applying data mining techniques and algorithms in various fields of human endeavour. However, as it is with almost every method, data mining also has some initial requirements in order for analysts to use it. Analyzed data should therefore be:

- complete (at least during analyzed time frame),
- consistent (significant changes during the analyzed time frame should not occur),
- correct (deprived of influence of human factors as much as possible).

There are several tasks which data mining is considered to be able to solve. We can aggregate those tasks into following groups:

a) Anomaly detection
  - detection of unwanted events,
  - treatment of outliers in data,

b) Association rules mining
  - discovering groups of products frequently bought together,
  - predicting possible defects,

c) Clustering
  - group technology coding,
  - analysis of stock items,

d) Regression
  - predicting demand,
  - modelling of indicator relationships,

e) Classification
  - decision support,
  - items sorting,

f) Summarization
  - visualization of performance,
  - discovering cyclical factors.

3.1.1 Data mining and its benefits for manufacturing planning and control

Planning and control of production process is a difficult task, because it is influenced by many factors which have impact on quality and time of delivery of product. Manufacturing control level employees are responsible for fulfilment of stated performance indicators. Every day they are forced to solve issues related to insufficient quality or performance at workplace. Employees or information systems monitor and record information about process states, which is later additionally discussed with manufacturing operators. Usual production feature is variability in performance and quality. This fact offers a question if there is a variant of manufacturing plan, in which planner/supervisor assigns manufacturing task to workplaces and operators in the way that he reaches the highest possible effectivity of production process.

2. Standard Data Mining Process

The most popular standard used by data mining experts is CRISP-DM. From a methodological point of view, it contains descriptions of main project stages, tasks contained in each of these stages, and relationships between these tasks. From a process model point of view, CRISP-DM offers an overview of data mining life-cycle, which consists of 6 stages, interconnected by arrows, as can be seen on Figure 2. These arrows symbolize the most important and the most frequent dependencies among individual stages. However, the order by which the stages are executed is not strict, and is often influenced by actual needs and constraints of the solution.

2.1 CRISP-DM stages

First stage, with which the data mining process should start, is a Business understanding. At the beginning it is crucial to establish goals, that data mining should reach in the company. Ambiguously defined goals can at the lead to receiving answers to wrong questions, which is a waste of valuable time and resources. In order to establish these goals properly, it is also necessary to evaluate a situation from various points of view, such as available resources, initial assumptions, constraints, risks, terminology being used and final costs and benefits of the solution. Final goals can therefore be considered as an output of this stage, usually along with project plan. [3]

In Data understanding stage, available data are being analyzed. This is also an important step, because insufficient understanding of data can lead to prolonging of following

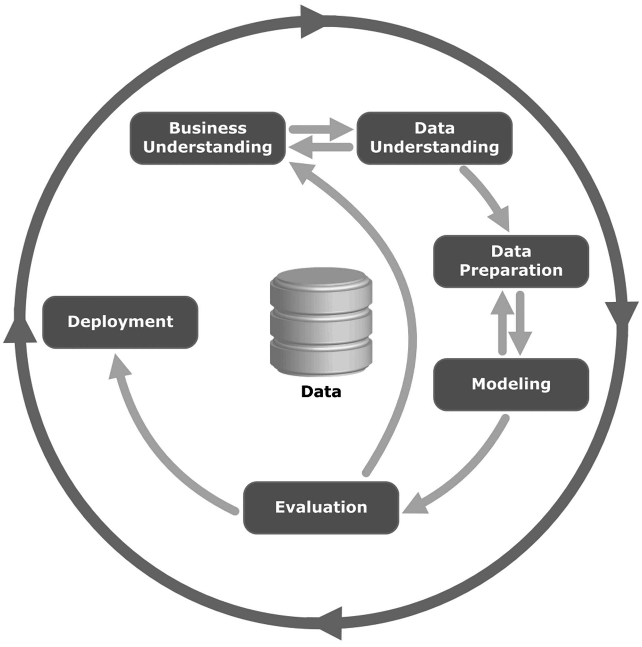

Fig. 2. CRISP-DM – data mining life-cycle

stage, where these data are being prepared for analysis. Initial data are being collected, described and experimented upon. At this stage, there also should be paid attention to verifying data quality.

It is noteworthy to mention, that Data preparation stage is usually the longest one in the overall data mining process, and sometimes can last more than 60 percent of time taken by one cycle. [11] At its beginning, data for processing are selected if they are likely to influence the solution. Selected data are cleaned, which can be interpreted as further selection of relevant subsets, insertion of default values, or mathematical approximation of missing values. This stage also includes insertion of computed attributes from available data, which often offer a more relevant representation of the system as in terms of modelling, which follows after this stage. Part of this stage is also an integration and type conversion.

At Modelling stage, data mining algorithms are applied to properly prepared data. Currently, there is a vast number of algorithms falling into a category of data mining, and so it is important to choose a proper subset of these algorithms, particularly that, which are capable to model the analyzed systems with smallest error. As a starting point, statistical properties of available data, and decision diagrams, such as scikit-learn algorithm cheat-sheet [6], can be used in order to choose an algorithm that fits the solution best. After selection of the most suitable algorithms, it is often necessary to set their initial parameters. Their application therefore requires at least some theoretical level of understanding of their functioning. Result of this stage is model with initial values of input parameters, which models the system best. These models are in fact a formalized knowledge, from which the company can benefit.

Evaluation stage is tied to results of final model, but also depends on goals defined in the first stage of business understanding, which are further dependant on organizational goals of the company, in which the data mining is taking place. Output of the evaluation is revision of process, as well as establishing further steps and decisions.

Deployment is a final stage, in which discovered knowledge is applied in the company environment. This can be done in a complex way, by designing and implementing an expert system, or simply, by setting, or establishing a way of setting process parameters in an company environment. Chosen way of deployment is a subject of Deployment plan and its realization is further monitored by the standards defined in Monitoring plan. At the end of this stage, final report and revision of the project along with documentation and discovered knowledge should be created as an aid for a point of another data mining cycle. Obviously, Deployment stage should be entered only after Evaluation stage returned plausible results. If it did not, it is often necessary to return back to the first stage, and redefine goals in an effort to better understand the analyzed system.

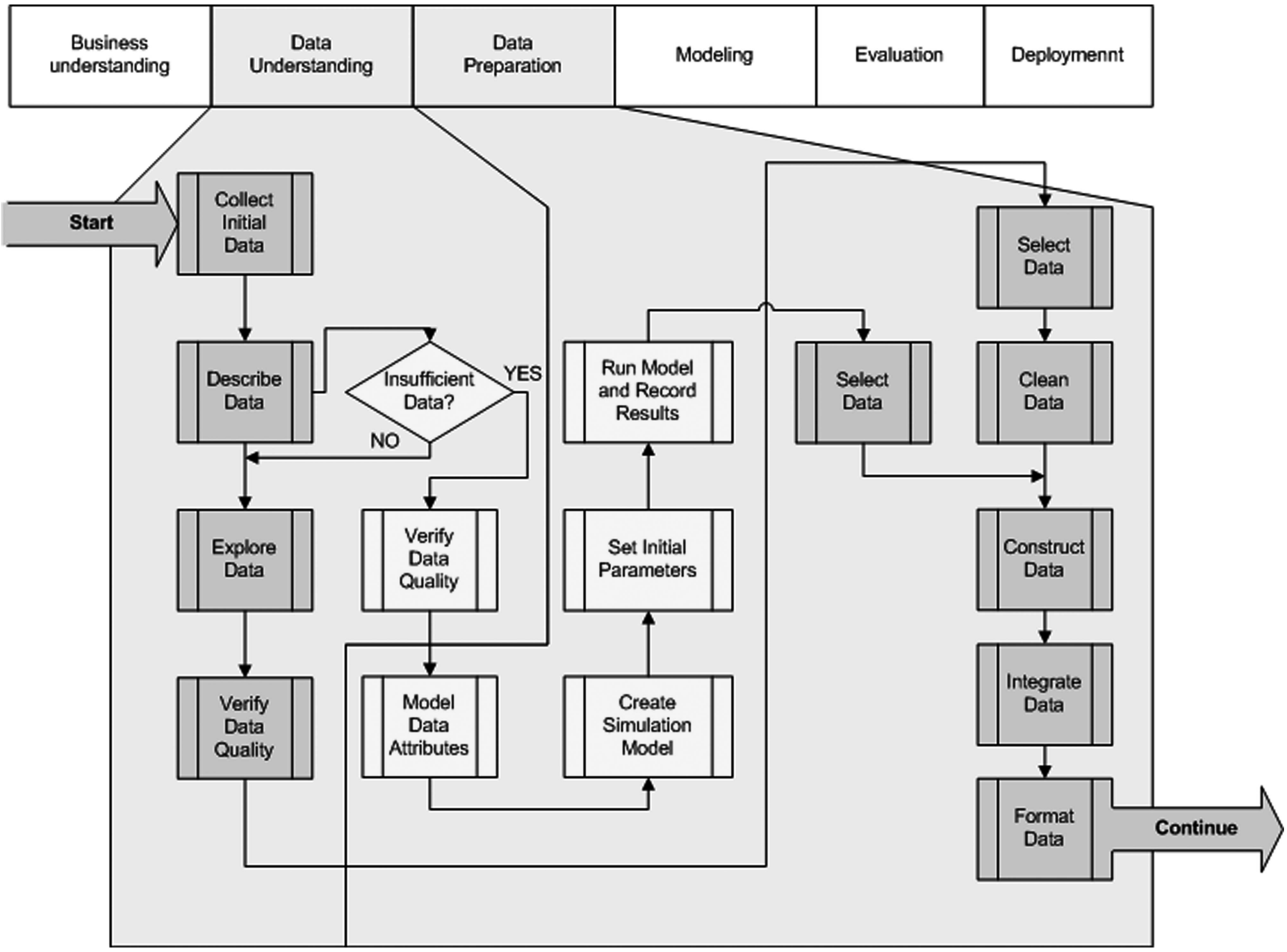

Fig. 3. Integrating simulation into data mining process

3. Integrating Simulation into Data Mining Process

A need to integrate simulation into application of data mining often arises from insufficient amount of relevant data, which could be used to model a behaviour of system in question. In spite of the fact, that companies gather vast volumes of data every day and the amount of this data has still an increasing tendency, their relevance in terms of data mining is often debatable. One of the most significant places where this phenomenon occurs are production processes, which are influenced by huge amount of external factors with variable characteristics. It is often very hard, and sometimes even impossible to find a sufficient amount of data, which represent enough of possible states of the system subjected to the analysis. The reason can sometimes be found in an insufficient data collection, but in some cases it is impossible to collect enough data due to a long processing period during which all the external factors could be taken into consideration and included into analysis. In this case, it is possible to use simulation in order to acquire sufficient amount of relevant data. Some examples of integrating simulation into data mining are optimization of batch size in push production system, where simulated production system generated data, on which it was possible to apply data mining algorithms in KNIME environment [1], optimization of number of kanban cards, where the simulated production system was a pull production system [2] and optimal selection of job which should be scheduled first in order to target job shop scheduling problem, where the simulation model consists of scheduler capable to dispatch individual jobs to machines with respect to their availability and various constraints derived, for instance, from bills of materials.

3.1 Data understanding

In Figure 3, referring to CRISP-DM standard, it is sensible to think about integrating simulation as early as in Data understanding stage and the final decision should emerge from Data description report, which reveals basic surface parameters, such as type and volume of records and data fields from available data sources.

If at this stage a conclusion is made, that the amount of available data is not sufficient, it is possible to ask a question, if the simulation has a chance to improve the results. However, even after decision for simulation is made, it is wrong to neglect verification of data quality, because parameters of simulation model are set based on it.

3.1 Data preparation

If the simulation approach is chosen, it is necessary to specify probability distributions of attributes, which would specify simulation model. Afterwards, it is possible to design the simulation model and set initial parameters, such as number of runs and range of analyzed input parameters i.e. independent variables to the problem. Then it is possible to run the simulation and record results, which would later on become a source for a further data mining analysis. These data do not require cleaning anymore, as the simulation model is not influenced by human factor, and thus its results do not contain erroneous or missing values, which would require further approximation. At this stage, it is also beneficial to construct, integrate and convert data to desired format, if it could lead to more accuracy of models at next stage. This depends on characteristics of simulation model, but also on very simulation environment, which defines the level into which we can reformat output data. Rest of the analysis can then continue with respect to CRISP-DM standard.

4. Conclusion

In this article, a way to integrate simulation into data mining has been described, particularly with respect to CRISP-DM standard. Recognition of possibility to apply simulation during the knowledge discovery process can in many cases, that occur in production environment, help to cross barriers that are laid by insufficient amount of relevant data, and thus make further data mining and subsequent profiting from discovered knowledge possible in more cases.

References:

[1] Horak F., Bubeník P., *Optimalizácia veľkosti výrobnej dávky v tlakovom výrobnom systéme aplikáciou získaných znalostí = (Batch size optimization in push production system by application of acquired knowledge)*. “Transfer inovácií” [elektronický zdroj]: internetový časopis o inováciách v priemysle. Č. 26, Available Online: http://www.sjf.tuke.sk/transferinovacii/pages/archiv/transfer/26-2013/pdf/182-185.pdf.

[2] Horak F., Bubeník P., *Získanie znalostí z ťahového výrobného systému prostredníctvom metód dolovania dát s využitím simulácie*. Průmyslové inženýrství 2013: mezinárodní studentská vědecká konference 3-4 října 2013, Valtice: sborník příspěvků. – Plzeň: SmartMotion, 2013, s. 33-36.

[3] *IBM SPSS Modeler CRISP-DM Guide*. 2011, Available Online: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf.

[4] Ismail R., *Data Mining In Production Planning and Scheduling: A Review*. 2nd Conference on Data Mining and Optimization 27-28 October 2009, Selangor, Malaysia. Available Online: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5341895.

[5] Kadav A., Kawale J., Mitra P., *Data Mining Standards*. DataMiningGrid, 2003. Available Online: http://www.datamininggrid.org/wdat/works/att/standard01.content.08439.pdf.

[6] Mueller A., *Machine Learning Cheat Sheet (for scikit-learn)*. 2013. Available Online: http://peekaboo-vision.blogspot.sk/2013/01/machine-learning-cheat-sheet-for-scikit.html.

[7] O’kane J., Spenceley J., Taylor R., *Simulation as an essential tool for advanced manufacturing technology problems*.” Journal of Materials Processing Technology” 2000, No. 107, pp. 412-424.

[8] Pyle D., *Systems, Simulation and Data Mining*. 2003, Available Online: http://www.iseesystems.com/community/connector/Zine/july-august_2003/pyle.html.

[9] Robinson S., *The Application of Computer Simulation in Manufacturing*. "Integrated Manufacturing Systems", Vol. 4, Iss: 4, pp. 18-23. Available Online: http://www.emeraldinsight.com/doi/abs/10.1108/09576069310044628.

[10] Sharma M., *Data Mining: A Literature Survey*. "International Journal of Emerging Research in Management & Technology", 2014, Vol. 3, No. 2, Available Online: http://www.ermt.net/docs/papers/Volume_3/2_February2014/V3N2-121.pdf.

[11] De Ville B., *Microsoft Data Mining: Integrated Business Intelligence for E-Commerce and Knowledge Management*. Digital Press, 2001.

Key words:
data mining, simulation, CRISP-DM

Abstract:
This article describes a way to integrate a simulation into a data mining technology, particularly with respect to CRISP-DM standard. Aim of this approach is to enable data mining in various cases, when available data do not meet all the requirements for data mining analysis. Solution is primarily tied to manufacturing companies environment, where there are many processes, that can be simulated, and thus the acquisition of sufficient volume of data for further analysis is possible.

ZINTEGROWANA SYMULACJA W EKSPLORACJI DANYCH

Słowa kluczowe:
eksploracja danych, symulacja, CRISP-DM

Streszczenie:
W artykule opisano sposób integracji oprogramowania symulacyjnego z technologią eksploracji danych, ze szczególnym uwzględnieniem standardu CRISP-DM. Celem takiego podejścia jest pozyskanie danych w przypadkach, gdy dostępne dane nie spełniają wszystkich wymagań związanych z analizą w systemie eksploracji danych. Zaproponowane rozwiązanie jest przede wszystkim związane z praktyką produkcyjną, gdzie realizowanych jest wiele procesów, które można komputerowo zasymulować, a tym samym można pozyskać wystarczające ilości danych do dalszych analiz.

Doc. Ing. Peter BUBENÍK, PhD.
ing. Filip HORÁK
Žilinská univerzita v Žiline
Strojnícka Fakulta
Katedra priemyselného inžinierstva
peter.bubenik@fstroj.uniza.sk
filip.horak@fstroj.uniza.sk