Data Pre-processing from Production Processes for Analysis in Automotive Industry

Veronika Simoncicova, Lukas Hrcka, Ondrej Tadanai, Pavol Tanuska, Pavel Vazan
Slovak University of Technology, Faculty of Materials Science and Technology in Trnava,
Institute of Applied Informatics, Automation and Mechatronics,
917 25 Trnava, J: Bottu 25, Slovak Republic
e-mail: veronika.simoncicova@stuba.sk, lukas.hrcka@stuba.sk,
ondrej.tadanai@stuba.sk, pavol.tanuska@stuba.sk, pavel.vazan@stuba.sk

Abstract. *Data pre-processing is an important part of the data mining process, because quality of the data used for an analysis mirrors in the results. If we have high-quality input data, then we will have high-quality results. The representation and quality of the instance data is very important, as analysing data not carefully screened for problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The article is focused on work with real data obtained from automotive industry related to breakdowns from the car body work. Data contain several problems with normalization and inconsistency. The presented analysis is provided as a partial result of the research and will serve to further investigation in the problem area. Our aim is design of function for modification data and acquiring new important information from this data. Results will be used for the future research.*

Keywords. RapidMiner; quality, data pre-processing, analysis; data; information

1 Introduction

Data pre-processing describes any type of processing performed on raw data with the purpose to prepare it for another processing procedure. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues.

Data pre-processing prepares raw data for further processing. Commonly used as a preliminary data mining practice, data pre-processing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

There is a number of different tools and methods used for pre-processing, including (Rouse, 2005):

- Sampling - selects a representative subset from a large population of data;
- Transformation - manipulates raw data to produce a single input;
- Denoising - removes noise from data;
- Normalization - organizes data for more efficient access;
- Feature extraction - pulls out specified data significant in a particular context.

Data goes through a series of steps during preprocessing (Technopedia):

- Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
- Data Integration: Data with different representations are put together and conflicts within the data are resolved.
- Data Transformation: Data is normalized, aggregated and generalized.
- Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.
- Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

Our research is based on production data export to CSV files. The goal of our research is to execute production data examination and further analysis in automotive industry using RapidMiner, an open source tool useful for data mining.

The automotive industry is more data-driven today than at any time in its history. In-car sensors, GPS tracking, automated manufacturing processes, and more are producing vast volumes of data that need to be analysed and understood. RapidMiner's Predictive Analytics platform enables car makers to derive value from this data by extracting the information hidden online, inside the vehicle, or at the plant with the purpose of better understanding product usage, preferences, and manufacturing processes to ensure quality and customer satisfaction.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both and to achieve better competitiveness. Users have to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large databases (Palace, 1996).

The pre-processing step is significant to solve several types of problems including noisy data, redundant data, missing data values, etc. All the inductive learning algorithms primarily rely on the product of this stage, which is the final training set. By selecting relevant instances, experts can usually remove irrelevant ones as well as noise and/or redundant data. Feature subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. The high-quality data will lead to high-quality results and to reduced costs for data mining. In addition, when a data set is too huge, it may not be possible to run a machine learning algorithm (Kotsiantis, 2007).

Data quality is defined in terms of accuracy, completeness, consistency, timeliness, believability, and interpretability. These qualities are assessed based on the intended use of the data. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Data cleaning is usually performed as an iterative two-step process consisting of discrepancy detection and data transformation (Han and Kamber 2006)

Knowledge discovery of data mining is a general process which in our research consists of several steps shown in Fig. 1:

1. Obtaining data and exporting dataq in required format: The first step in this process is to create an input file containing a list of attributes and values. Subsequently, the modified data is better processed and the work is efficient and more effective.
2. Pre-processing: CSV file is used as the input data source for RapidMiner. This source data is loaded into RapidMiner and functions for eliminating inconsistent or erroneous data are used. Such modified data are ready to be processed with analytical techniques.
3. Result: The output from the process consists of pre-processing data interpreted as partial results for the company and consequently used as a basis data set for future analysis.

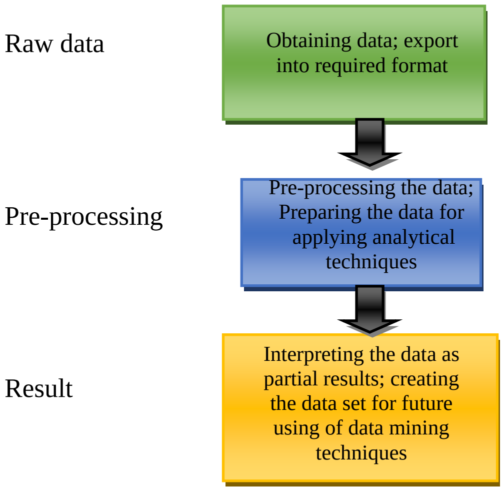

Raw data
Obtaining data; export into required format
Pre-processing
Pre-processing the data; Preparing the data for applying analytical techniques
Result
Interpreting the data as partial results; creating the data set for future using of data mining techniques

Figure 1. The steps of knowledge discovery

2 Design Pre-processing Using RapidMiner

2.1 RapidMiner

Currently, rapid miner is one of the most used open source predictive analytics platforms utilised for data analysis. It is accessible as a stand-alone application for information investigation and as a data mining engine for the integration into own products. Rapid miner provides an integrated environment for data mining and machine learning procedures, including (Grupta and Malhotra, 2015):

- extracting the data from various source systems; transforming the data and loading into a data warehouse (DW) or data repository other applications,
- data pre-processing and visualization,
- predictive analytics and statistical modelling, evaluation, and deployment.

Providing learning schemes, models and algorithms from WEKA and R scripts makes the system even more powerful (Grupta and Malhotra, 2015).

Rapid miner provides a graphical user interface (GUI) used to design and execute analytical workflows. Those workflows form a process, which consists of multiple Operators. GUI allows connecting the operators with each other in the process view. Each independent operator carries out one task within the process and forms the input to another operator in the workflow. The major function of a process is the analysis of the data which is retrieved at the beginning of the process (Grupta and Malhotra, 2015).

Rapid miner offers a large amount of different operators extensible using available extensions. There are packages for text processing, web mining, WEKA extensions, R scripting, series extension, python scripting, anomaly detection and more (Akthar and Hahne).

2.2 Data Pre-processing

The first step is data export from Legato system into CSV format of files, where the data from car body work is stored. Legato is system for saving data from production, for example messages, breakdowns etc. The data contains breakdowns and messages from manufacturing. Breakdowns are automatically recorded in the Legato system. Our aim is to partially prepare the data, to clean the data and to unify the data written for the first analysis.

In Fig. 2, you can see process of loading data using "Read CSV" operator. This operator is used to read CSV files. Read CSV is an abbreviation for Comma-Separated Values. CSV files have all values of an example in one line. Values for different attributes are separated by a constant separator. It may have many rows. Each row uses a constant separator for separating attribute values. CSV name suggests that the attribute values would be separated by commas, but other separators can also be used. "Append" operator joins the two data files in one file on the base of the same attributes. This operator builds a merged example set from two or more compatible example sets by adding all examples into a combined set.

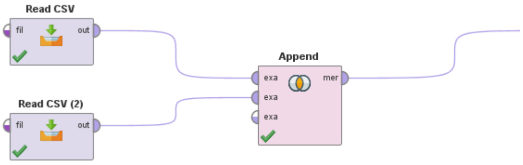
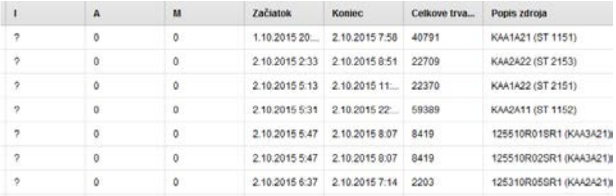


Figure 2. Joining of the two data files

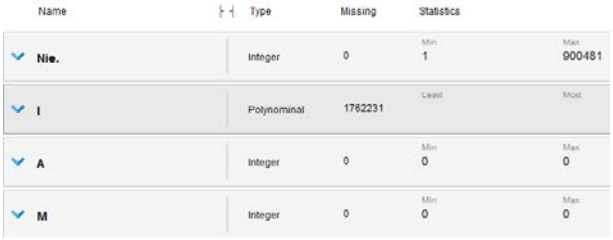
CSV file contains data from car body work and involves records concerning breakdowns on each machine,including exact time, duration of breakdowns, group of machines, date and brief descripton of breakdowns. Data for change are depicted in Fig. 2. The problems of data set cover noisy data, inconsistent and redundant data, missing data values and wrong type of data.



I	A	M	Začiatok	Koniec	Celkove trva...	Popis zdroja
?	0	0	1.10.2015 20:...	2.10.2015 7:58	40791	KAA1A21 (ST 1151)
?	0	0	2.10.2015 2:33	2.10.2015 8:51	22709	KAA2A22 (ST 2153)
?	0	0	2.10.2015 5:13	2.10.2015 11:...	22370	KAA1A22 (ST 2151)
?	0	0	2.10.2015 5:31	2.10.2015 22:...	59389	KAA2A11 (ST 1152)
?	0	0	2.10.2015 5:47	2.10.2015 8:07	8419	125510R01SR1 (KAA3A21)
?	0	0	2.10.2015 5:47	2.10.2015 8:07	8419	125510R02SR1 (KAA3A21)
?	0	0	2.10.2015 6:37	2.10.2015 7:14	2203	125310R05SR1 (KAA2A21)

Figure 3. Example set data from car body work

Performing the first analysis (Fig. 3), we acquired information about missing values, unused attributes selected from our dataset, for example "Nie", "I", "A", "M". These attributes contain zero value or missing value and therefore, they are not applicable for the analysis.



Name	Type	Missing	Statistics	
Nie.	Integer	0	Min 1	Max 900481
I	Polynominal	1762231	Least	Most
A	Integer	0	Min 0	Max 0
M	Integer	0	Min 0	Max 0

Figure 4. First Analysis

Fig. 4 portrays the complete process model of the first data pre-processing representing the operators needed for preparing data.

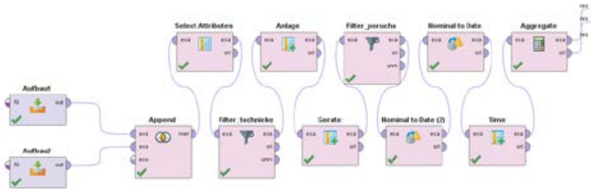
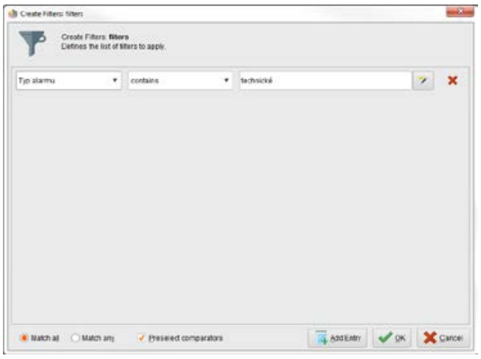


Figure 5. Final model of data pre-processing

Operator „Select Attributes" selects which attributes of an example set should be kept and which attributes should be removed. This feature is used in cases when not all attributes of an Example set are required; it helps to select the required attributes.

Next step is to select rows, which shall be analysed. "Filter_technicke" and „Filter_porucha" (Fig. 6) operators were used to select which examples (i.e. rows) of the set should be kept and which examples should be removed. Examples satisfying the given condition are kept, remaining examples are removed.



Create Filters: filters

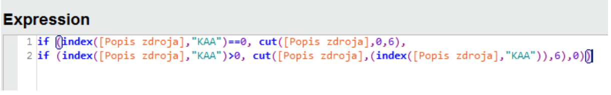
Create Filters: **filters**
Defines the list of filters to apply.

Typ alarmu | contains | technické

● Match all ○ Match any ✓ Preseled comparators

Figure 6. Filters

The used records are not standardized, which represents a serious problem due to the data inconsistency. Data are incosistent in several attributes, therefore, we created new attribute „Anlage" containing exact information on the placement of a group of robots, because this information was available scattered into several attributes. Attribute „Anlage“ is generated using „Generate attribute“ operator, containing the expression (Fig. 7): if (index([Popis zdroja],"KAA")==0, cut([Popis zdroja],0,6), if (index([Popis zdroja],"KAA")>0, cut([Popis zdroja],(index([Popis zdroja],"KAA")),6),0)). This expression index begining of character „KAA“. This expression indexes items starting with the character string “KAA”. After indexing the suitable items, the results are compared with the condition and expression is cut to the number of characters specified according to our needs.

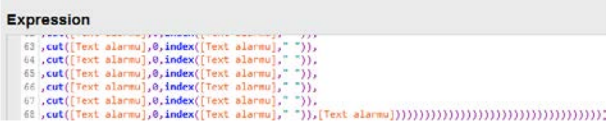


Expression

```
if ((index([Popis zdroja],"KAA")==0, cut([Popis zdroja],0,6),
if (index([Popis zdroja],"KAA")>0, cut([Popis zdroja],(index([Popis zdroja],"KAA")),6),0)()
```

Figure 7. Expression 1

In the next step, the new attribute „Gerate“ was created, containing names of individual machines. The attribute „Gerate“ was generated using „Generate attribute“ operator, which contains the expression (Fig. 8): “if (starts([Text alarmu],"AS")==true,cut([Text alarmu],0,index([Text alarmu]," ")), if (starts([Text alarmu],"BA")==true,cut([Text alarmu],0,index([Text alarmu]," ")), if (starts([Text alarmu],"AE")==true,cut([Text alarmu],0,index([Text alarmu]," ")), if (starts([Text alarmu],"Ku")==true,cut([Text alarmu],0,index([Text alarmu]," ")), if (starts([Text alarmu],"LM")==true,cut([Text alarmu],0,index([Text alarmu]," "))………, if (starts([Text alarmu],"12")==true,cut([Text alarmu])))))))))))))”. This expression works in similar manner as the previous expression, however, we used the function „start“ to compare the specificed characters; if condition is true, function „cut“ cuts the name of the machine up to the indexed space; otherwise the “alarm text” is used.

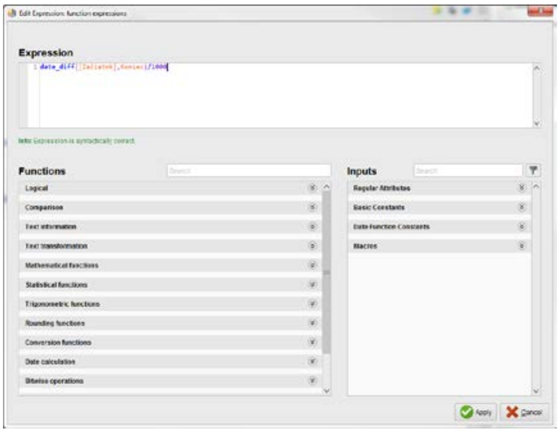


Expression

```
,cut([Text alarmu],0,index([Text alarmu]," ")),
,cut([Text alarmu],0,index([Text alarmu]," ")),
,cut([Text alarmu],0,index([Text alarmu]," ")),
,cut([Text alarmu],0,index([Text alarmu]," ")),
,cut([Text alarmu],0,index([Text alarmu]," ")),
,cut([Text alarmu],0,index([Text alarmu]," ")),[Text alarmu])))))))))))))))))))))))))))))))))))
```

Figure 8. Expression 2

Inadequate type of „Celkové trvanie“ attribute represents a further problem area. The received data on the length of breakdown duration was not exported correctly, as the attribute of type “nominal” was changed numerical data causing loss of all items with duration longer than one tousand seconds. The format of date was necessary to change using the function „date_diff“ (Fig. 9). The function deducts the first and the last specifed date and returns exact duration of individual breakdowns.



Edit Expression: function expressions

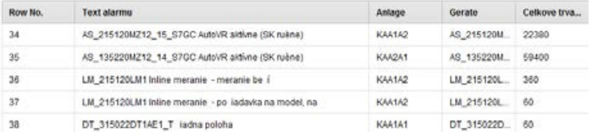
Expression

```
date_diff([Zaciatok],[Koniec])/1000
```

Info: Expression is syntactically correct.

Figure 9. Expression 3

All changes will contribute to better, faster and more efficient analysis of data. Modifying the data is very important part of data pre-processing, as the quality of acquired results depends on the quality of the used data. In Fig. 10 data after all changes are depicted.



Row No.	Text alarmu	Anlage	Gerate	Celkove trva...
34	AS_215120MZ12_15_S7GC AutoVR aktivne (SK ruène)	KAA1A2	AS_215120M...	22380
35	AS_135220MZ12_14_S7GC AutoVR aktivne (SK ruène)	KAA2A1	AS_135220M...	59400
36	LM_215120LM1 Inline meranie - meranie be í	KAA1A2	LM_215120L...	360
37	LM_215120LM1 Inline meranie - po iadavka na model, na	KAA1A2	LM_215120L...	60
38	DT_315022DT1AE1_T iadna poloha	KAA1A1	DT_315022D...	60

Figure 10. Data after changes

As a final procedure of data pre-processing, utilising the “Agregate” operator and using a new attribute „Anlage“, we aquired the breakdown counts for individual groups of machines (Fig. 11).

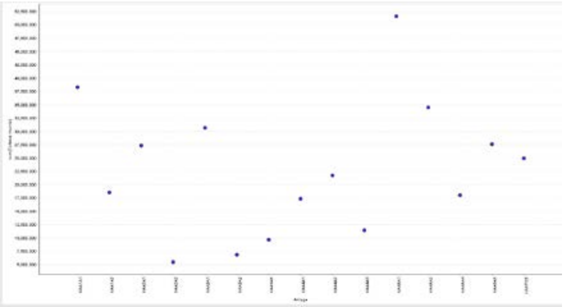
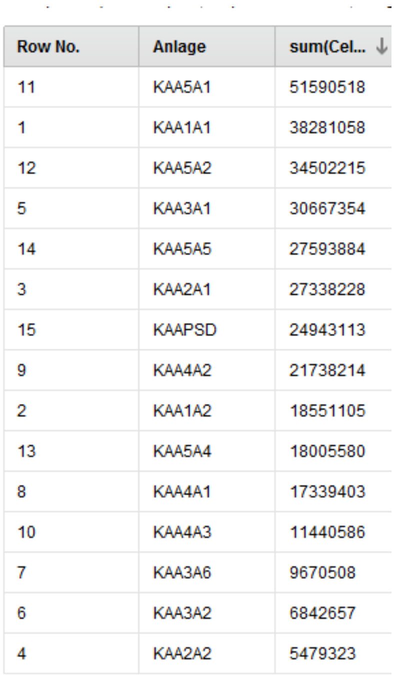


Figure 11. Graph of count breakdowns

Based on the conducted analysis of the problem area, we concluded that the group identified as KAA5A1 (Fig. 11) is responsible for the major part of all breakdowns in the analysed production system, showing the highest breakdown rate for the period of six months. KAA5A1 is the area of machines, where belong riveting machines, cutting machines and other robots.

Total number of breakdowns (Fig. 12) per machine is 51,590,518, in comparison to the second group of machines, the increase reaches over 34 percent.



Row No.	Anlage	sum(Cel... ↓
11	KAA5A1	51590518
1	KAA1A1	38281058
12	KAA5A2	34502215
5	KAA3A1	30667354
14	KAA5A5	27593884
3	KAA2A1	27338228
15	KAAPSD	24943113
9	KAA4A2	21738214
2	KAA1A2	18551105
13	KAA5A4	18005580
8	KAA4A1	17339403
10	KAA4A3	11440586
7	KAA3A6	9670508
6	KAA3A2	6842657
4	KAA2A2	5479323

Figure 12. Count of breakdowns

3 Evaluation and results

This paper aims to introduce this new area of data pre-processing as a critical step in a data mining project as well as its practical part using Rapid miner in order to show the effect of data pre-processing.

In our research, data pre-processing was applied to data obtained from the field of automobile industry. The sample data used for the analysis were from the period of 10/01/2015 to 03/30/2016. At the beginning of the process, the data was necessary to edit because the data was incomplete and inconsistent. Using the tool RapidMiner, we modified the data and we unified and removed unneeded attributes.

Applying the analysis, we found out, that the group of machines named by a KAA5A1 has the worst results for the period of last six months.

The resulting processed data provides a partial result for the company and also the data serves as a basic data set for the further research using advanced data mining techniques.

4 Acknowledgment

This publication is the result of implementation of the project VEGA 1/0673/15: “Knowledge discovery for hierarchical control of technological and production processes” supported by the VEGA. This publication is the result of implementation of the project: "UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU - CAMBO" (ITMS: 26220220179) supported by the Research & Development Operational Program funded by the EFRR.

The authors would like to thank for financial contribution from the STU Grant scheme for Support of Young Researchers.

5 References

Palace, B., (1996). What is datamining? Retrieved from: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

Kotsiantis, S. B., (2007). Data Preprocessing for Supervised Leaning, Retrieved from: http://waset.org/publications/14136/data-preprocessing-for-supervised-leaning

Rouse, M., (2005). Data pre-processing, Retrieved from: http://searchsqlserver.techtarget.com/definition/data-preprocessing

Technopedia, Data preprocessing, Retrieved from: https://www.techopedia.com/definition/14650/data-preprocessing

RapidMiner, Today´s automotive markets must move beyond traditional strategies Internet, Retrieved from: https://rapidminer.com/industry/automotive/

Han, J., and Kamber, M., (2006) *Data Mining: Concepts and Techniques* (Second Edi.). San Francisco: Elsevier Inc.

Grupta, G., and Malhotra, S., (2015) “Text Documents Tokenization for Word Frequency Count using Rapid Miner (Taking Resume as an Example)”, International Journal of Computer Applications (0975-8887), International Conference on Advancement in Engineering and Technology (ICAET 2015)

Akthar, F., and Hahne, C., “RapidMiner 5 Operator Reference”.