Analysis and structuring diagnostic large volume data of technical condition of complex equipment in transport

V V Vychuzhanin\(^1\), N R Rudnichenko\(^1\), Z Sagova\(^2\), M Smieszek\(^3\), V V Cherniavskyi\(^4\), A I Golovan\(^5\) and M V Volodarets\(^6\)

\(^1\) ONPU Information Technologies Department, Shevchenko avenue 1, 65044 Odessa, Ukraine
\(^2\) University of Žilina, Faculty of Mechanical Engineering, Department of Automation and Production Systems, Univerzitná 1, Žilina, 010 26, Slovakia
\(^3\) Rzeszow University of Technology, Department of Quantitative Methods, 35-959 Rzeszow, al. Powstancow Warszawy 8, Poland
\(^4\) Kherson State Maritime Academy, Marine Engineering Faculty, 20 Ushakova avenue, Kherson, 73000, Ukraine
\(^5\) Odessa National Maritime University, 34 Mechnikova, Odessa, 65029, Ukraine
\(^6\) Ukrainian State University of Railway Transportation, Ukraine

E-mail: msemieszk@prz.edu.pl

Abstract. The paper presents the results of the classification analysis model for structuring of processed large volumes of heterogeneous diagnostic data about the technical state of complex equipment in transport development and research. Concept for the description and structuring of big data is proposed based on the formation of a metadata scheme using logical breakdown of all technical diagnostic data on the output variable - the technical condition of complex technical equipment in transport. A functional assessment of the technical condition complex technical system’s elements in transport is developed based on the application of methods for assessing structural and functional risks of failures. The article presents the results of assessing the accuracy of the input data sets classification using created decision trees models to effectively structuring and presenting the data in order to ensure that the procedures for their further analysis are performed. As a result of using the developed simulation model of structuring and presenting large heterogeneous diagnostic data volumes about the state of complex technical equipment in transport the time costs were reduced and the efficiency of analytical operations to study data for solving diagnostic problems and predicting complex system’s technical condition was improved.

1. Introduction
Currently, there is a steady trend of active development modern technical systems software and hardware infrastructure on various transport types [1–3]. This is largely due to the increased requirements for providing the necessary level of vehicles operational processes automation; navigation support; technical system failures diagnostics; formalization of external influences from the environment; the exchange of information between onboard systems through the use of wireless technologies and devices [4–8]. Such features affect on the growth of the general level transport means informatization [9–11]. This necessitates effective information measuring systems usage, means and channels of communication integration, which allow to get and to transfer technical systems elements...
numerical values of the functional parameters in real time [12]. In this regard, the number of data sources is increasing, which significantly affects the amount of transmitted, processed and stored information of different types [13].

In the case when the volumes of generated data exceed tens and hundreds of gigabytes, it is advisable to use approaches and technologies for processing big data volumes (Big Data), since classical methods, algorithms and database management systems (DBMS) are not always able to effectively carry out the necessary transaction processes [14–16].

2. Description of problem

Currently, solutions based on the Map Reduce model of distributed computing are being actively used to support the processing and analysis of Big Data, which allows providing parallel processing of data and increasing the processing speed [17–19]. Due to the use of distributed file system (HDFS), flexible organization of processing and data convolution to the control server becomes possible. The advantages of this approach are the automation of the computational nodes distribution to clusters within a computer network, which allows using the computational capabilities of an unlimited number of hosts, implementing data backup algorithms to ensure the reliability of their storage, as well as supporting software implementations for most modern high-level programming languages [20–22].

Significant disadvantages of this model are the complexity of streaming data processing in real time, difficulty in deploying the system if there is an unstable data transmission channel with low bandwidth, a resource-intensive procedure for visualizing data by iterations of their processing, as well as in the case of a short online transactions large number. These features are inherent to transport modern technical systems. Therefore, the implementation of this model is not always justified and effective, it often requires appropriate modification [23].

It is known that increasing the length of SQL commands for aggregated operations of searching, sampling, inserting and saving data in relational database management systems (DBMS) increases the complexity of generating transactional queries to databases (DB).

Depending on the hardware resources used, this can significantly reduce the speed and efficiency of data processing in real time. Therefore, the use of non-relational databases (NoSQL), such as MongoDB, Redis, HBase, Firebase, and others [24] is appropriate in practice as Big Data storages. The advantages of using NoSQL in ensuring the processing of Big Data are due to factors: a high level of flexibility in providing the necessary level of data scaling; no restrictions on stored data types; support of a key-value data representation which is convenient for processing; document based approach.

In contrast to financial and social data, large volumes of technical information are characterized by static and stable characteristics, since a significant part of the time of transport technical system (TTS) operation under normal conditions its parameters do not change dramatically. Such specificity necessitates the constant reorganization of data in order to eliminate low-informative arrays and use statistical analytical procedures for generalization and aggregation of meaningful information, which requires additional computational resources and operations [25]. This is required for a more economical use of production capacity of computing facilities and reducing the data analysis processes time spent [26].

To solve such problems, it is advisable to develop an analytical unit that implements parsing and data classification in order to compress them as a separate node, which is convenient for integration in implementing micro-service architecture.

An important task in the processing of large amounts of data generated during the components operation (sensors for removing dynamic parameters of system elements, data receiving and transmission modules, control modules and others) of a modern vehicle is effective segmentation and clustering of information into separate fragments (sets) based on using hierarchical and non-hierarchical algorithms [27].

The advantage of this approach is the ability to perform more accurate data analysis by parallel use of different types of models for each selected set, which allows the comparison and analysis procedures for evaluation its adequacy, as well as helps to reduce the number of computational operations performed on processing a whole array of data.
Actual direction and purpose of this work is to study the possibilities of structuring and a short presentation of incoming TTS data to provide opportunities for their further processing and analysis. The obtained models can be integrated into decision support systems based on the use of artificial intelligence (AI) methods to extract new useful knowledge from the collected data and identify hidden patterns (Data mining) [28]. Such systems will reduce the time spent on a number of analytical operations, allowing automatically evaluate various operational scenarios of TTS usage, their subsystems and functional components states, build analytical models and forecast the technical condition of the system within a reasonable planning horizon.

The efficiency of the organization of the Data mining process for Big Data essentially depends on the type of data presentation, their consistency, the quality of cleaning, integrity, orderliness and detail. Currently, there are a significant number of approaches in AI that allow the design of learning systems to automate the processes of searching, accumulating, structuring data and knowledge [29]. The most promising of these are: artificial immune and neural networks, deep machine learning and hybrid adaptive learning algorithms, which are based on the principles and models of the human brain functioning. Such methods allow speeding up the process of analyzing big data significantly, but their application requires the initial structuring of data and the identification of significant features for the models input and output data formation [30]. Within the framework of the problem it is advisable to use a hierarchical approach in the formation of the data structure, and therefore it is possible to use the decision tree method (DCT).

3. General concept
Existing methods for constructing DCT can be used to describe key ontological meta-information about heterogeneous data sets of large volumes, with the aim of providing a more compact form of their presentation with subsequent processing and analysis [31]. DCT can be visualized by mapping a connected oriented acyclic graph [32, 33].

The branches of the graph of the constructed DCT can store the values of the attributes, which are the functional parameters of the analyzed TTS elements, which the objective function is depended of, and its value is displayed on the DCT’s leaves.

The technical parameters of the TTS elements are received and converted from the corresponding sensors via wired and wireless communication channels, which is held under the onboard data acquisition system controlling.

The obtained data is aggregated, structured as belonging to the subsystems and stored in *.json files, after that they are sent in separate threads to the cloud-based IAAS infrastructure using the virtual working environment and containerization using Docker tools. The data integrated within the IAAS system is processed by the consolidation module according to specified criteria (compactness, information content, filtering) for their further consistent and uniform distribution within the individual data stores (DS) as nodes of the data space (pool). The data presentation and structuring module (DPSM) makes a request to retrieve the required data of a given volume into the pool, in response to which a set of data from the corresponding DS is returned.

As a result of DPSM operation, DCT models were obtained in the form of an ordered metadata scheme in json and xml formats, which are sent for further storage in distributed repositories that operate under the control of the HDFS file system. Further data mining of stored meta-information is possible through the use of analytical software (Knime, SPSS Statistics, QIWare, etc.) or own system.

A ship power plant based on the Wartsila 16V50DF was chosen as a technical transport system for research [34].

The implementation of the stage is the result of performing the complexes of operations of representation and structuring (figure 1).

The main parameters governing the logic of the process of constructing the DPSM models of DCT are:
- $CR$, tree separation criterion is an index of heterogeneity;
- $DP$, tree maximum depth level;
- **CN**, value of the confidence level of building a model for estimating the pessimistic error of cutting off tree branches;
- **LT**, minimum value of a leaf of a tree;
- **LS**, minimum sample size in the training subset;
- **PP**, number of alternative tree nodes for an early stop DCT building.

**Figure 1.** General algorithm for constructing a DCT model of big data descriptions.

The following indicators are used for all TTS elements as functional parameters (attributes) for building a DCT model:
- structural risk of system element failure ($R_{str}$), representing the level of its vulnerability in the TTS topology, it takes a real data type in the range [0,1], a dimensionless quantity;
- functional risk of system element failure ($R_{func}$), reflects the level of its vulnerability in the dynamics of the functioning of the TTS, it takes a real data type in the range [0,1], a dimensionless quantity;
operation mode of the system element (OM), characterizes the intensity and load of its use, it takes a string data type in the form of two possible options «Regular» and «Irregular»;

- duration of the system element operation (OD), determines the period of its functional use within the system, accepts an integer data type in the range [1,20], measured in years;

- system element maintainability degree (MD), formalizes the level of an element's exposure to functional modifications in case of failure in order to operatively ensure its further functioning, accepts a string data type in the form of three possible options «Low», «Middle» and «High»;

- system element physical deterioration degree (W), displays the level of element’s damage, affecting its target functioning, accepts an integer data type as a percentage from 0 to 100%;

- regular time between failures of the system element (MTBF), characterizes the average time between occurrences of failures in his work, takes an integer data type in the range from 0 to 1,000,000, measured in hours;

- number of element repairs completed (RN), describes the history of the technical measures to restore its performance, takes an integer data type in the range [1,10], a dimensionless quantity;

- average market value of the item (EC), characterizes the criticality of element’s failure from a financial point of view and the potential cost of replacement, accepts an integer data type in the range [1000,100000], measured in arbitrary units;

- element performance (PR), accepts a string data type in the form of three possible options «Low», «Middle» and «High».

The integral indicator of the element's assessment (the target variable for classification) of a system is its technical condition (TC), the output value of a DCT leaf can be one of two classes: «Acceptable» and «Unacceptable».

The final data functionality for a DCT based description is:

\[ TS' = \left\{ \text{Acceptable, Unacceptable} \right\} \]

\[ TS' = \left\{ R_\text{str}, R_\text{func}, OM, OD, MD, W, MTBF, RN, EC, PR \right\} \]  \hspace{1cm} (1)

\( R_\text{str} \) and \( R_\text{func} \) values are formed on the basis of the building cognitive-imitation models for assessing the risk of TTS elements failures approach and using the damaging modeling impulses and normalizing effects proposed in the works [1, 12, 35]. The efficiency and degree of maintainability are evaluated by experts based on the use of successive comparisons method, taking into account the Spearman’s rank correlation coefficient.

To study the possibility of increasing the efficiency of solving the classification problem in constructing the DCT model, it is proposed to use the following heterogeneity indices: information entropy (IG); information gain (GR); gini index (GI); standard accuracy (AC).

When constructing a DCT model, it is necessary to carry out a numerical evaluation of the accuracy of the classification carried out to analyze the adequacy of its work. In this regard, in order to assess the quality of building a model, it is advisable to use and implement the following numerical metrics: \( ACC \) – relative number of properly classified examples in the data sample; \( CE \) – relative number of misclassified examples in a data sample; \( KPP \) – kappa-statistics, allows for the accounting of random correct classification; \( WMR \) – weighted average completeness classification; \( WMP \) – weighted average classification accuracy.

4. Structuring model development

In order to test the proposed concept of the ontological data meta-information scheme structural representation the process of functioning of the DPSM was modeled by constructing DCT using the designated TC TTS functional using means of the cross-platform software Rapid Miner Studio. A fragment of the results of statistical processing of one of the presented data samples to form a DCT model is shown in figure 2. The imported data from the Retrieve block is transferred to the Set Role block, which is used for setting logical roles to the necessary attributes of the analyzed data sample. Using this block, the target output variable is selected to form the DCT structure, which is the technical state of the TTS element in our case.
The configured data set is fed to the input of the Split Data block, which breaks the sample into subsets (training and test), in accordance with the selected ratio. The decision tree block implements operators for specifying all previously defined parameters and constraints when building a DCT model, allowing selection and configuration of various combinations of structuring the processed data by TTS elements in order to identify the most informative presentation of them in graphical tree form. The determination of the numerical values of the designated metrics for assessing the quality of constructed DCT model work is carried out by using the Performance block.

![Figure 2. A fragment of the data presented statistical processing results for a DCT model forming.](image)

The block composition of the DCT simulation model is shown in figure 3.

The obtained DCT models allow us to structure and display the results of solving the problem of classifying data according to the degree of the TTS efficiency with varying degrees of detail and flexibility in interpretation. It is necessary to conduct a numerical evaluation of the used metrics for a comprehensive results analysis.
5. Analysis of the results

In the process of carrying out the DCT modeling process, the operations performed on constructing tree models and estimating the numerical parameters of metrics were logged into text files for further analysis. Summary results of the assessment of DCT model metrics for the GR, IG, GI and AC indices are given in Table 1. From the point of view of the convenience of presenting meta-information about data and its further interpretation, the DCT model implemented on the basis of GR is more detailed and flexible. Classification based on using GI and AC indices is less accurate (low values of calculated metrics).

| Indexes used | GR  | IG  | GI  | AC  |
|--------------|-----|-----|-----|-----|
| Metric values | ACC | CE  | KPP | WMR | WMP |
| ACC          | 98.01 % | 96.79 % | 91.24 % | 93 % |
| CE           | 1.99 % | 3.21 % | 8.76 % | 7 % |
| KPP          | 0.966 | 0.95 | 0.89 | 0.9 |
| WMR          | 98.48 % | 97.57 % | 92.91 % | 91.73 % |
| WMP          | 97.22 % | 95.70 % | 91.26 % | 89.95 % |

A fragment of the obtained logical rules according to the performed DCT model based on the GA data classification index:

StructRisk > 0.490: Low {High=0, Middle=0, Low=5} StructRisk ≤ 0.490
| FailureMean > 180000: Low {High=0, Middle=0, Low=2} |
| FailureMean ≤ 180000 |
| OperDurat > 3.500 |
| Wear > 7.500 |
| Average cost > 17500 |
| StructRisk > 0.435 |
| OperMode = Irregular: Low {High=0, Middle=0, Low=1} |
| OperMode = Regular: Middle {High=0, Middle=1, Low=0} |
| StructRisk ≤ 0.435: Middle {High=0, Middle=7, Low=0} |
| Average cost ≤ 17500: High {High=1, Middle=0, Low=0} |
| Wear ≤ 7.500: High {High=1, Middle=0, Low=0} |
| OperDurat ≤ 3.500: High {High=2, Middle=0, Low=0} |
Result attribute density distribution $R_{str}$ for acceptable and unacceptable TC TTS for averaged sample of data is shown in figure 4. It should be noted that the of the acceptable TC class values growth is observed in the range of 0.25–0.33 and is less broad in shape compared to the range of unacceptable class TC values of 0.31–0.41. This is evidence that the TC system is not acceptable and requires prompt intervention.

Figure 4. Structure risk attribute density distribution.

The scatterplot for identifying the correlation between the structural risk of failures and the technical state of the TTS elements is shown in figure 5. On the diagram of each record of the analyzed data set there corresponds a point, the coordinates of which correspond to the classified values of the structural

Figure 5. Scatter plots identify correlations between structural risk and technical state of elements.
risk along the abscissa axis and TC TTS along the ordinate axis. Analysis of the constructed scatterplot allows us to confirm the earlier statement about the prevalence of cases when TC is unacceptable. This visualization of the data makes it possible to additionally detect small outliers of values at the boundary ranges that can be cut off during further analysis.

The distribution density of data on the TC of the TTS, depending on the wear and repairs carried out on the elements of the technical system for assessing the impact on their TC in the form of a heat map is shown in figure 6.

The dots show the values of permissible and non-permissible TC of the TTS elements. The analysis of the formed regions allows us to assert a more pronounced correlation between the low performance and non-acceptable values of the TC of the TTS elements, and a less pronounced correlation in the case when the performance is medium and high.

Figure 6. Fragment of the heat map of the wear data and repairs distribution density on the elements of the technical system to assess the impact on their TC.

From the analysis of the DCT models and graphical diagrams, it should be noted that a higher level of correlation is observed between the parameters of the structural risk of TTC failures, the degree of technical deterioration and the regulated time between failures. The values of these parameters have the highest impact on the TC and level of performance of TTS elements, which determines the criticality in their operational monitoring.

The resulting DCT structure is further needed to speed up the process of analyzing data on the generated metadata arrays by reducing the computational cost of data processing and avoiding the need for sending and converting additional samples.

6. Conclusion
Analysis of the research results allowed us to identify the main advantages of using decision trees to represent and structure large amounts of data, which consist of:

- providing processing capabilities of heterogeneous (discrete, continuous) multi-class data sets, which are characterized by the presence of a significant number of features;
- no need for additional operations to transform, normalize and scale the values of features;
- elimination of missing and incorrect values of the used features support;
- implementation of additional metrics for assessing the significance of attributes;
- integration of the results with software solutions through the use of json and xml formats.
It should be noted the high total size of the constructed data representation models (about 30 megabytes per 20,000 records) can be reduced by applying logical convolution operations according to separate criteria, archiving algorithms and using gradient boosting by composing compositions from other existing machine learning algorithms.

Analysis of the created DCT models research results made it possible to establish that the most appropriate approach is to build a tree based on the use of the GR index.

The model is 8–10 % more efficient from the point of view of classification quality compared to other models, which determines its acceptability for solving problems of structuring large amounts of data and convenience for hierarchical visualization of the relationships between selected data segments.

The developed model can be used as a separate module of a decision support system that evaluates and predicts the technical condition of the complex technical transport systems elements in various conditions of their operation.

7. References
[1] Boyko V, Rudnichenko N, Kramskoy S, Hrechukha Y and Shibaeva N 2016 Concept implementation of decision support software for the risk management of complex technical system Advances in Intelligent Systems and Computing 512 255–269
[2] Gritsuk I, Zenkin E Y, Bulgakov N, Golovan A et al. 2018 The complex application of monitoring and express diagnosing for searching failures on common rail system units SAE Technical Paper 2018-01-1773
[3] Kuric I, Mateichyk V, Smieszek M, Tsiuman M et al. 2018 The peculiarities of monitoring road vehicle performance and environmental impact MATEC Web of Conferences 244 03003
[4] Kuric I, Tlach V, Ságová Z, Císař M et al. 2018 Measurement of industrial robot pose repeatability MATEC Web of Conferences 244 01015
[5] Gritsuk I, Gutarevych Y, Mateichyk V and Volkov V 2016 Improving the processes of preheating and heating after the vehicular engine start by using heating system with phase-transitional thermal accumulator SAE Technical Paper 2016-01-0204
[6] Klets D, Gritsuk I, Makovetskyi A, Bulgakov N et al. 2018 Information security risk management of vehicles SAE Technical Paper 2018-01-0015
[7] Gritsuk I, Volkov V, Mateichyk V, Grytsuk Y et al. 2018 Information model of V2I system of the vehicle technical condition remote monitoring and control in operation conditions SAE Technical Paper 2018-01-0024
[8] Hahanov V, Gharibi W, Litvinova E, Chumachenko S et al. 2017 Cloud-driven traffic monitoring and control based on smart virtual infrastructure SAE Technical Paper 2017-01-0092
[9] Gritsuk I, Volkov V, Mateichyk V, Gutarevych Y et al. 2017 The evaluation of vehicle fuel consumption and harmful emission using the heating system in a driving cycle SAE Int. J. Fuels Lubr 10 236–248
[10] Rabinovich E, Gritsuk I V, Zuiev V, Zenkin E Y et al. 2018 Evaluation of the powertrain condition based on the car acceleration and coasting data SAE Technical Paper 2018-01-1771
[11] Zaharchuk V, Gritsuk I V, Zaharchuk O, Golovan A et al. 2018 The choice of a rational type of fuel for technological vehicles SAE Technical Paper 2018-01-1759
[12] Vychuzhanin V, Rudnichenko N, Shybaiev D, Gritsuk I et al 2018 Cognitive model of the internal combustion engine SAE Technical Paper 2018-01-1738
[13] Rudnichenko N D, Vychuzhanin V V and Shybaiev D S 2017 The use of cluster data analysis for highlighting the measure of complex technical systems factors influencing the similarity performance Informatics and Mathematical Methods in Simulation 3 214–219
[14] Shybaiev D S, Vychuzhanin V V, Shibaeva N O and Rudnichenko N D 2017 Optimization of the selection and analysis of information in multi-structured data stores Informatics and Mathematical Methods in Simulation 3 318–324
[15] Shybaiev D S, Vychuzhanin V V, Shibaeva N O and Rudnichenko N D 2018 Optimization of methods for forecasting, processing and analyzing information in multi-structured data stores Informatics and Mathematical Methods in Simulation 1 78–85

[16] Shybaiev D S, Vychuzhanin V V and Rudnichenko N D 2018 Improving the efficiency of methods for the selection and analysis of information in the data-structured data warehouses, 21st International Youth Forum "Radio Electronics and Youth in the 21st Century" 5 221–222

[17] Vychuzhanin V V, Shibaev D S, Boyko V D, Shibaeva N O and Rudnichenko N D 2017 Big data mapping in the geopositioning systems for fishing industry International Scientific and Technical Conference on Computer Sciences and Information Technologies 28–31

[18] Chekharin E E 2016 Big data: big problems Perspectives on science and education 3 7–11

[19] Biktimirov M P, Elizarov A M and Shcherbakov A Y 2016 Trends in the development of big data processing technologies and tools for storing multi-format data and analytics Russian Digital Libraries Journal 5 390–406

[20] Hyunjoung L and Sohn I 2016 Fundamentals of Big Data Network Analysis for Research and Industry (Wiley)

[21] Li K-C, Jiang H and Zomaya A Y 2017 Big Data Management and Processing (CRC Press)

[22] Liu Sh, McGree J, Ge Z and Xie J 2016 Computational and Statistical Methods for Analysing Big Data with Applications (Elsevier)

[23] Maheshwari A 2015 Big Data Essentials (Anil Maheshwari)

[24] Moniruzzaman B M and Hossain S A 2013 NoSQL database: New era of databases for big data analytics - classification, characteristics and comparison International Journal of Database Theory and Application 6 1–14

[25] Pore S S and Pawar S B 2015 Comparative Study of SQL & NoSQL Databases International Journal of Advanced Research in Computer Engineering & Technology 4 1747–1753

[26] Ghavami P 2016 Big Data Governance: Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics (CreateSpace Independent Publishing Platform)

[27] Sedkaoui S 2018 Data Analytics and Big Data (Wiley)

[28] Mulyukova KV 2019 Comparative analysis of Data Mining modern tools Young scientist 1 19–21

[29] Shcherbyna A D, Shybaiev D S, Rudnichenko N D and Shibaeva N O 2018 Comparative analysis of existing areas in Data mining Project, Program, Portfolio Management The Third International Scientific-practical Conference 88–90

[30] Rokach L, Maimon O 2008 Data Mining with Decision Trees: Theory and Applications (World Scientific)

[31] Subbotin S A Hoffman E A 2011 The use of the apparatus of decision trees in the task of diagnosing car bodies System analysis. Computer Science. Management: II All-Ukrainian Scientific and Practical Conference 211–212

[32] Shamaev D R 2017 Decision trees for tasks of constructing a rating of commercial banks Technical sciences: problems and perspectives: materials V International. scientific conference 18–22

[33] Druzhnov P N 2013 Parallel implementation of the gradient-boosting algorithm for decision trees Computational Methods and Programming: New Computational Technologies 2 109–114

[34] Wartsila, “Wartsila50DF - high output with fuel flexibility” 2018 Available from: https://www.wartsila.com/products/marine-oil-gas/engines-generating-sets/dual-fuel-engines/wartsila-50df

[35] Vychuzhanin V, Rudnichenko N, Boyko V, Shibaeva N and Konovalov S 2016 Devising a method for the estimation and prediction of technical condition of ship complex systems Eastern-European Journal of Enterprise Technologies 84 4–11