Article citation info:
KOZIELSKI M, SIKORA M, WRÓBEL Ł. Decision support and maintenance system for natural hazards, processes and equipment monitoring. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2016; 18 (2): 218–228, http://dx.doi.org/10.17531/ein.2016.2.9.

Michał KOZIELSKI
Marek SIKORA
Łukasz WRÓBEL

DECISION SUPPORT AND MAINTENANCE SYSTEM FOR NATURAL HAZARDS, PROCESSES AND EQUIPMENT MONITORING

SYSTEM WSPOMAGANIA DECYZJI DLA MONITOROWANIA ZAGROŻEŃ NATURALNYCH, PROCESÓW I URZĄDZEŃ

This paper presents the DISESOR integrated decision support system and its applications. The system integrates data from different monitoring and dispatching systems and contains such modules as data preparation and cleaning, analytical, prediction and expert system. Architecture of the system is presented in the paper and a special focus is put on the presentation of two issues: data integration and cleaning, and creation of prediction model. The work contains also two case studies presenting the examples of the system application.

***Keywords**: decision support system, prediction, expert system, data cleaning, process monitoring, device monitoring, hazard.*

W pracy przedstawiono zintegrowany system wspomagania decyzji DISESOR oraz jego zastosowania. System pozwala na integrację danych pochodzących z różnych systemów monitorowania i systemów dyspozytorskich. Struktura systemu DISESOR składa się z modułów realizujących: przygotowanie i czyszczenie danych, analizę danych, zadania predykcyjne oraz zadania systemu ekspertowego. W pracy przedstawiono architekturę systemu DISESOR, a szczególny nacisk został położony na zagadnienia związane z integracją i czyszczeniem danych oraz tworzeniem modeli predykcyjnych. Działanie systemu przedstawione zostało na dwóch przykładach analizy dla danych rzeczywistych.

***Słowa kluczowe**: system wspomagania decyzji, czyszczenie danych, predykcja, system ekspertowy, monitorowanie procesów, monitorowanie urządzeń, monitorowanie zagrożeń.*

1. Introduction

Coal mining is a heavy industry that plays an important role on an energy market and employs hundreds of thousands of people. Coal mining is also an industry, where large amount of data is produced but little is done to utilise them in further analysis. Besides, there is a justified need to integrate different aspects of coal mine operation in order to maintain continuity of mining what can be done by introduction of a decision support system (DSS).

Currently coal mines are well equipped with the monitoring, supervising and dispatching systems connected with machines, devices and transport facilities. Additionally, there are the systems for monitoring natural hazards (methane-, seismic- and fire hazards) operating in the coal mines. All these systems are provided by many different companies, what causes problems with quality, integration and proper interpretation of the collected data. Another issue is that the collected data are used chiefly for current (temporary) visualisation on boards which display certain places in the mine. Whereas, application of domain knowledge and the results of historical data analysis can improve the operator's and supervisor's work significantly.

For example, due to the short-term prognoses about methane concentration, linked with the information about the location and work intensity of the cutter loader, it is possible to prevent emergency energy shutdowns and maintain continuity of mining (the research on this methodology was discussed in [27]). This will enable to increase the production volume and to reduce the wear of electrical elements whose exploitation time depends on the number of switch-ons and switch-offs.

It is possible to see the rising awareness of monitoring systems suppliers who has started to understand the necessity to make the next step in these systems development. Therefore, the companies providing monitoring systems seek their competitive advantage in equipping their systems with knowledge engineering, modelling and data analysis methods. This is a strong motivation to consider a DSS presented in this paper.

The goal of this paper is to present an architecture of the DISESOR integrated decision support system. The system integrates data from different monitoring systems and contains an expert system module, that can utilise domain expert knowledge, and analytical module, that can be applied to diagnosis of the processes and devices and to prediction of natural hazards. Special focus of the paper is put on the data integration and data cleaning issues, such as outlier detection, realised by means of the data warehouse and the ETL process. The work also contains a more detailed presentation of the prediction module and two case studies showing real applications of the system.

The contribution of the paper consists of the architecture of the DISESOR integrated decision support system, its data repository and prediction module. Additionally, it covers the presentation of the issues connected with the preparation and cleaning of the data collected by monitoring systems, especially outlier detection. Finally, the contribution covers case studies presenting application of the described system to abyssal mining pump stations diagnostics and methane concentration prediction in a coal mine.

The structure of the paper is as follows. Section 2 presents the works related to the presented topic. The architecture of the DISESOR system and its data repository are presented in section 3. The more detailed descriptions of the data preparation and cleaning and prediction modules are presented in sections 4 and 5 respectively. The case studies of abyssal mining pump stations diagnostics and methane concentration prediction task are presented in sections 6 and 7 respectively. Section 8 presents the final conclusions.

2. Related work

The typical environments deployed in a coal mine are monitoring and dispatching systems. These systems collect a large number of data which can be utilised in further analysis, e.g., on-line prediction of the sensor measurements, which area was surveyed in [11]. Such analysis can address different aspects of coal mine operation such as, e.g., equipment failure or natural hazards.

The examples of the research in the field of natural hazards in an underground coal mine cover, e.g., methane concentration prediction and seismic hazard analysis. The research on the prediction of the methane concentrations was presented in [26, 27, 28]. Application of data clustering techniques to seismic hazard assessment was presented in [15]. There are also approaches to prediction of seismic tremors by means of artificial neural networks [8] and rule-based systems [9]. Each research listed above is a standalone approach not incorporated into any integrated system.

Analytical methods that were mentioned require the data which are extracted, cleaned, transformed and integrated. Decision support systems utilise a data repository of some kind, e.g., a data warehouse [13]. The critical dependence of the decision support system on a data warehouse implementation and an impact of the data quality on decision support is discussed in [17].

There are applications of machine learning methods to diagnostics of mining equipment and machinery presented in literature [4, 5, 12, 19, 30]. The issue of mining industry devices diagnostics was raised among others in the works [7, 10, 18, 25, 31]. Besides, some initial concepts of the system that processes data streams delivered by the monitoring systems were presented in [6].

However, to the best of the authors knowledge there is no example of the integrated decision support system for monitoring processes, devices and hazards in a coal mine (except the work dealing with DSS for coal transportation [14], which loosely corresponds to the given topic).

3. System architecture

The general architecture of the DISESOR integrated decision support system is presented in Fig. 1. The architecture of the system consists of: Data repository, Data preparation and cleaning module, Prediction module (that are presented in more detail in the following sections), Analytical module and Expert system module (shortly presented below, as they are not the main focus of the paper).

3.1. Decision support system

The core of analytical, prediction and expert system modules is based on the RapidMiner [22] platform. The RapidMiner environment was customised to the requirements of the non-advanced user by disabling unnecessary options and views. Therefore, an advanced user can use the whole functionality of RapidMiner, whereas the non-advanced user can use such thematic operators as e.g., “Solve a methane concentration prediction issue” or “Solve a seismic hazard issue”. Additionally, due to the target application of the system in Polish coal mines the RapidMiner environment was translated into Polish (for this reason several figures in this work contain Polish names). Finally, RapidMiner was extended in the created application by additional operators wrapping R [21] and MOA (Massive On-line Analysis) [1] environments.

The goal of the Data preparation and cleaning module, which is referred further as ETL2, is to integrate the data stored in data warehouse and process them to the form acceptable by the methods creating prediction and classification models. In other words the ETL2 module prepares the training sets.

Prediction module is aimed to perform incremental (on-line) learning of predictive models or apply classification and prediction models created in analytical module for a given time horizon and frequency of the values measured by the chosen sensors. This module also tracks the trends in the incoming measurements. The created predictive models are adapted to the analysed process on the basis of the incoming data stream and the models learnt on historical data (within the analytical module). The module provides the interfaces that enable the choice of quality indices and their thresholds that ensure the minimal prediction quality. If the quality of predictions meets the conditions set by a user, the predictions will be treated as the values provided by a soft sensor. They can be further utilised by e.g., expert system but also they can be presented to a dispatcher of a monitoring system.

Expert system module is aimed to perform on-line and off-line diagnosis of machines and other technical equipment. It is also aimed to supervise the processes and to support the dispatcher or expert decision-making with respect to both technical condition of the equipment and improper execution of the process. The inference process is performed by means of classical inference based on stringent rules and facts, fuzzy inference system or probabilistic inference based on belief networks. Additionally, the system contains a knowledge base editor that allows a user to define such rules and networks.

Analytical module is aimed to perform analysis of historical data (off-line) and to report the identified significant dependencies and trends. The results generated by this module are stored in the repository only when accepted by a user. Therefore, this module supports a user in decision-making of what is interesting from monitoring and

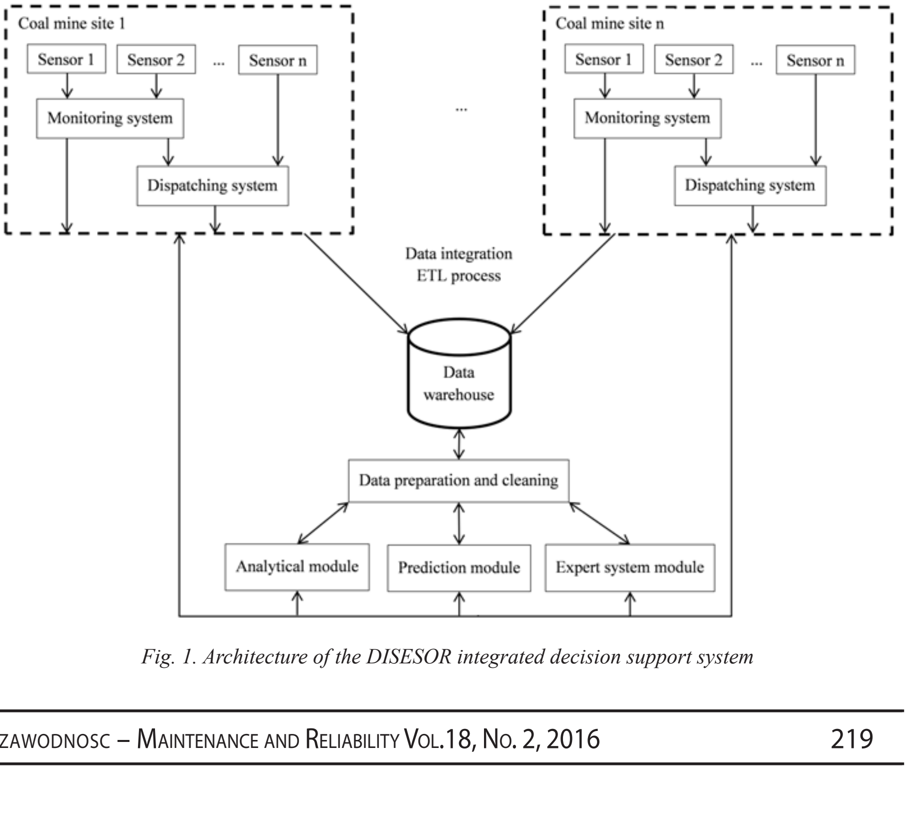


Fig. 1. Architecture of the DISESOR integrated decision support system

prediction point of view. Besides, it provides additional information that can be utilised to enrich the knowledge of expert system or that can be utilised to comparative analysis. The module supports identification of changes and trends in the monitored processes and tools and it also enables to compare the operator's and dispatcher's work.

3.2. Data repository

Data repository was designed as a data warehouse of a snowflake structure, that is presented in a general form in Fig. 2. The structure of a data warehouse results from the analysis of databases of the existing monitoring systems and the characteristics of the known sensors. The full list of tables with their description is presented in Table 1.

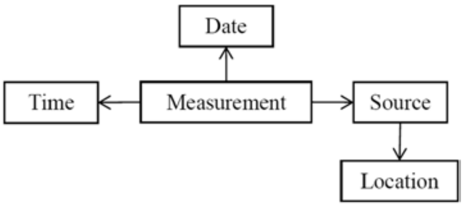


Fig. 2. Simplified schema of data repository

Table 1. Tables creating a data warehouse structure

Measurement	Value of a measurement
State	State of a measurement, e.g., *alarm*, *calibration*, *breakdown*
Discretisation	The measured values can be of discrete type
Time	Time of a measurement, range [00:00:00, 23:59:59], 1 second resolution
Time_category	Category, e.g., *mining* or *no mining*
Date	Date of a measurement
Location	Location of the measurement source
Location_attribute	Characteristics of the given location
Location_hierarchy	Hierarchical structure of location
Source	Measurement source, e.g., sensor or device
Source_attribute	Characteristics of the given source

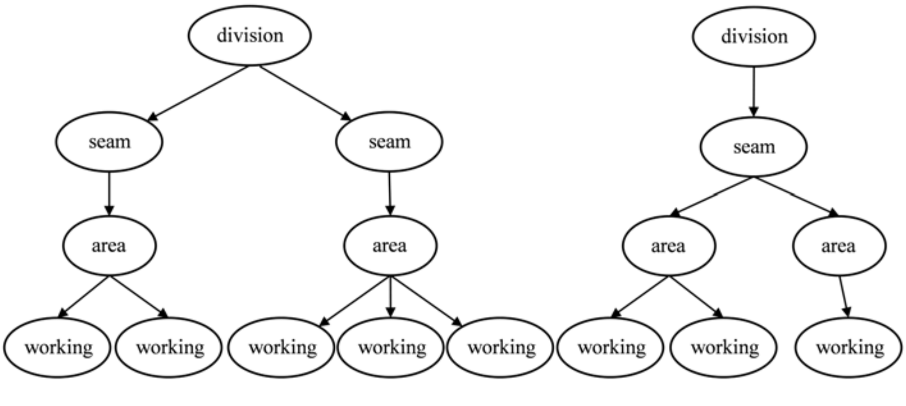


Fig. 3. Location hierarchy in a coal mine

The central table of the data repository is *Measurement* where all the measurements are stored. The dimensions related to the *Measurement* table are *Date*, *Time* and *Source*. *Date* and *Time* describe when the measurement was registered, whereas *Source* describes what registered the given measurement. The *Source* table contains among others such information about sensors/devices as:

- name (e.g. MM256),
- description (e.g. methane meter number 256),
- type name (e.g. methane meter),
- measured quantity (e.g. methane concentration),
- measurement unit (e.g. %CH4),
- name of a system that collects the data (e.g. THOR),
- range of measurements.

The *Source* table is described by means of *Location* dimension, that describes where in a coal mine it is located. The location has hierarchical structure, some sample hierarchy is presented in Fig. 3. The top-most level of the hierarchy is formed by coal mine divisions. Divisions consist of seams, which are divided into mining areas. At the bottom of the hierarchy there are mining workings.

The data warehouse is loaded with data by means of the ETL process designed for the main monitoring and dispatching systems for coal mining, which are deployed in Poland, Ukraine and China, e.g., THOR dispatching system [24] or Hestia natural hazards assessment system [9]. The ETL process was designed by means of Open Talend Studio [29].

During the tests of the created solution the data warehouse was loaded with 800 million records what resulted in 200 GB of data. It enabled the performance tests and optimisation of both the logical data warehouse structure and database management system (PostgreSQL [20]). As a result the *Measurement* data table was partitioned according to the months of measurements and the indices for foreign keys in this table were created. On the DBMS side several configuration parameters were adjusted, e.g., shared_buffers, work_mem, checkpoint_segments, effective_cache_size.

4. Data preparation and cleaning

The goal of ETL2 (Data cleaning and preparation) module is to deliver integrated data (in a form of a uniform data set) coming from chosen sources (especially sensors) in a chosen time range. Therefore, in this section the issues of frequency adjustment, aggregation and missing values imputation are presented. The outlier detection issues are extended in the subsection 4.1.

Measurements can be collected with different frequencies. Additionally, some systems collect a new measurement only after significant (defined in a monitoring system) change of the measured value. Table 2 presents how the measurements of two methanometers can look like, when collected directly from the data warehouse. The ETL2 process uniforms the data to the form, where each recorded measurement represents the time period defined by a user, e.g., 1 second (Table 3).

Within the ETL2 module there are also executed procedures of data cleaning, that identify outlier values and impute the missing values. These tasks are realised both by means of the simple functions presented below and by means of operators available in RapidMiner environment.

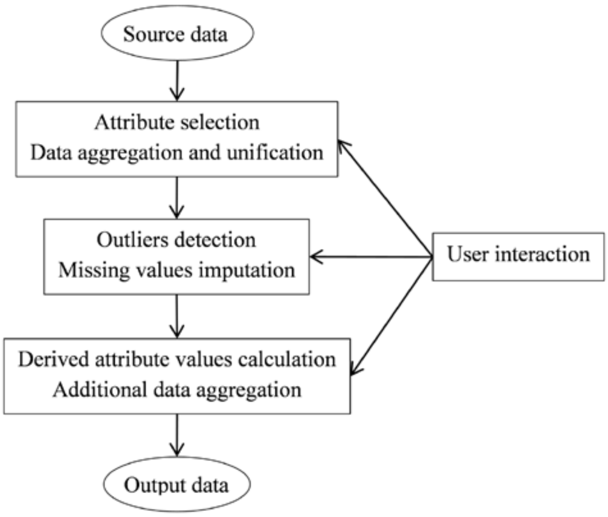
Another operations performed by means of the methods included in the ETL2 module are data aggregation (e.g., 10 measurements are replaced with 1 measurement) and manually performed definition of derived variables (e.g., a new variable can be calculated as a sum of the values of two other variables). The general scheme of data processing within ETL2 module is presented in Fig. 4.

Table 2. Data collected directly from data warehouse (- means that the measurement value does not change, ? means a missing value)

MN234 [%CH$_4$]	MN345 [%CH$_4$]	T [s]
0.1	0.1	0
0.2	-	1
-	0.2	4
0.5	?	7
0.3	0.3	9

Table 3. Data prepared to the further transformation, cleaning, etc.

MN234 [%CH$_4$]	MN345 [%CH$_4$]	T [s]
0.1	0.1	0
0.2	0.1	1
0.2	0.1	2
0.2	0.1	3
0.2	0.2	4
0.2	0.2	5
0.2	0.2	6
0.5	?	7
0.5	?	8
0.3	0.3	9


Source data

Attribute selection
Data aggregation and unification

Outliers detection
Missing values imputation

User interaction

Derived attribute values calculation
Additional data aggregation

Output data

Fig. 4. General characteristics of the data processing in ETL2 module

All the phases of processing presented in Fig. 4 are performed as separate RapidMiner operators. As a result of the processing performed by means of the ETL2 module we receive a data set that can be either analysed (by means of analytical module), or utilised to prediction model creation (by means of prediction module), or utilised within diagnostic process (by means of expert system module).

In order to select the variables that should be analysed a user can utilise THOR dispatching system, where each sensor (and attributes) are presented on a map of the region of interest. An exemplary screen of the THOR system is presented in Fig. 5. The system that is being created enables in turn, data (time-series) visualisation in order to select the time periods, that are the most interesting from the analyst point of view. Fig. 6 presents the visualisation of time-series consisting of several thousands of records. The developed operator creating such visualisation utilises R environment [21].

Aggregation of the measurements replaces several values with a single one. The period of aggregation is chosen by a user, who sets a number of measurements that should be aggregated or a time unit defining the windows containing measurements to be aggregated. The following aggregation operators are available for each attribute: average, minimum, maximum, median, dominant, the number of occurrences. For each record being the result of the aggregation there is calculated a weight, that is inversely proportional to the number of missing values existing in the aggregated data. The weight calculation is also based on a weighted average for all the attributes. This approach enables us to reduce the number of missing values in data and introduce weights that can be utilised by the chosen methods (e.g. rule induction).

The operator that imputes missing values performs the analysis of each attribute separately. The following methods that change the value or imputing the missing value can be utilised:

- a logical expression defining the replacing values (e.g. replace each value <1 with „low state"),
- the way how to receive the replacing values:
  - the value set by a user,
  - the last valid measurement,
  - average of the neighbouring measurements (with the parameter defining the number of neighbours),
  - linear regression of the two points (the last one before missing values section and the first one after this section),
  - linear regression of the data preceding missing values (with the parameter defining the window size).

The maximal number of consecutive missing values that can be imputed is defined as a separate parameter, as imputing the values for the long breaks in the measurements has no practical meaning. If the

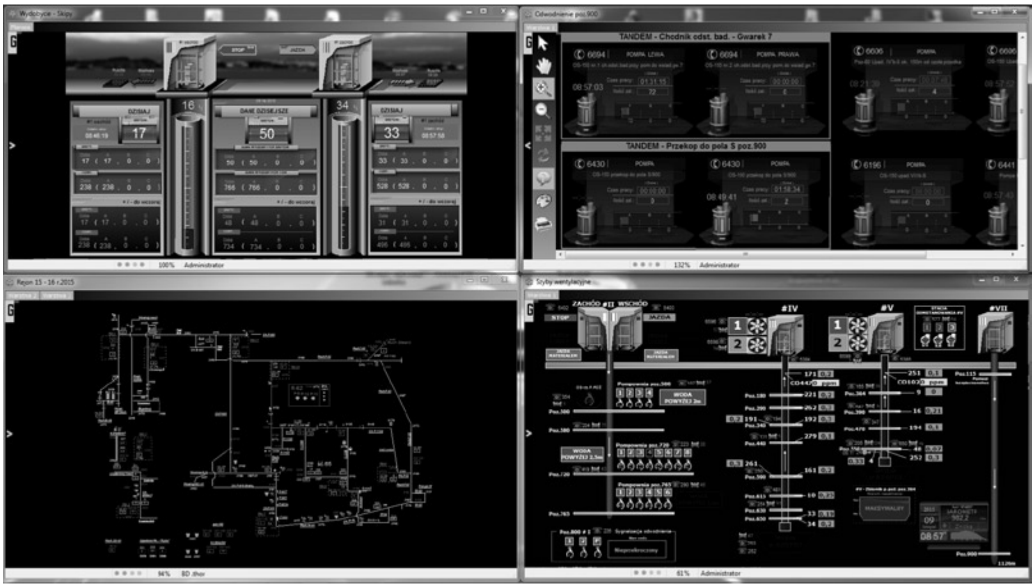

TANDEM - Chodnik odst. bad. - Gwarek 7
TANDEM - Przekop do pola 5 poz.900
DZISIAJ
DANE DZISIEJSZE
ZACHÓD #II WSCHÓD
#IV
#V
#VII

Fig. 5. Visualisation available in THOR dispatching system presenting a topology of the sensors

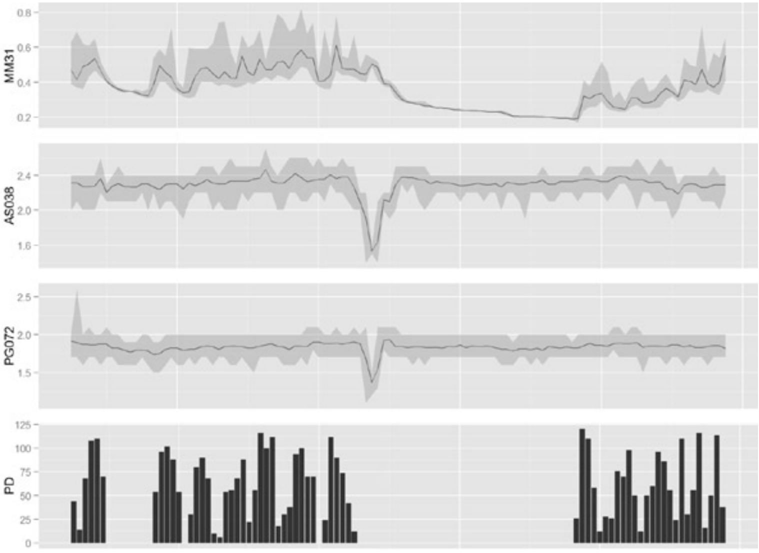


Fig. 6. Visualisation of exemplary time-series: methane concentration, air flow and mining cycle on a chosen longwall

resulting data set still contains missing values, the analyst can use a number of methods that are able to analyse data with missing values.

Introduction of a new derived variable can cover, among others, introduction of delays (the values of the previous measurements) or calculation of increments and trends (e.g. as an ordinal - increases, decreases). Another operator enables data smoothing by means of different filters (e.g. average, median). Finally, the last operator enables creation of dependent variable (decision variable). Typically, this variable contains the moved forward values of the chosen attribute, what enables to receive a proper prediction horizon. The operator defining the dependent variable has expanded functionality what enables e.g. to define the dependent variable as a maximal value of a given attribute in a defined time interval (e.g. 3 to 6 minutes in advance).

It is also important that within the developed framework the operators can be applied multiple times and in unrestricted order. Moreover, it is possible to pre-process data by means of the operators delivered by RapidMiner, that are dedicated to multidimensional analysis/ identification of outliers and missing values (e.g. the operator applying local k-NN to missing values imputation).

When data preprocessing is finished, the whole process is saved according to XML-based RapidMiner standard, that was created for the needs of the system. Thereby, the prediction module and expert system module are able to transform the incoming data to the form that is acceptable by prediction and inference solutions. The incoming data in this case are collected on-line directly from the monitoring systems.

4.1.Outlier detection methods

Analysis of data coming from several underground coal mines showed that the missing values are relatively rare because most of the monitoring systems are the safety ones, where undisturbed data transfer is of the high importance. The methods that are based on linear interpolation or the last measured value approach fit well to the imputation of missing value task. Among 800 million data measurements that were loaded to DISESOR data repository only 0.5% contained missing values. These missing values consisted of single missing measurements, tens of missing measurements or longer periods of missing values being a result of transmission break (in this last case there is no effective method of missing value imputation).

The issue that is much more complex is detection of outlier values that can be a result of measurement interference. RapidMiner environment, except manual (expert) elimination of missing values that were mentioned above, offers several methods of automatic outlier detection. Such analysis is multidimensional, what means that impact of each variable of a given record is verified. Four methods of this type were evaluated during this research. These methods are characterised by high effectiveness in outlier detection and efficiency, as they do not require extensive computations. The methods that were chosen are the following [22]:

- Detect Outliers – Density (CDODe) – the method identifying the outliers on the basis of their density. The method requires two parameters. A record is identified as an outlier if there is at least the defined ratio of other records (where the ratio is given

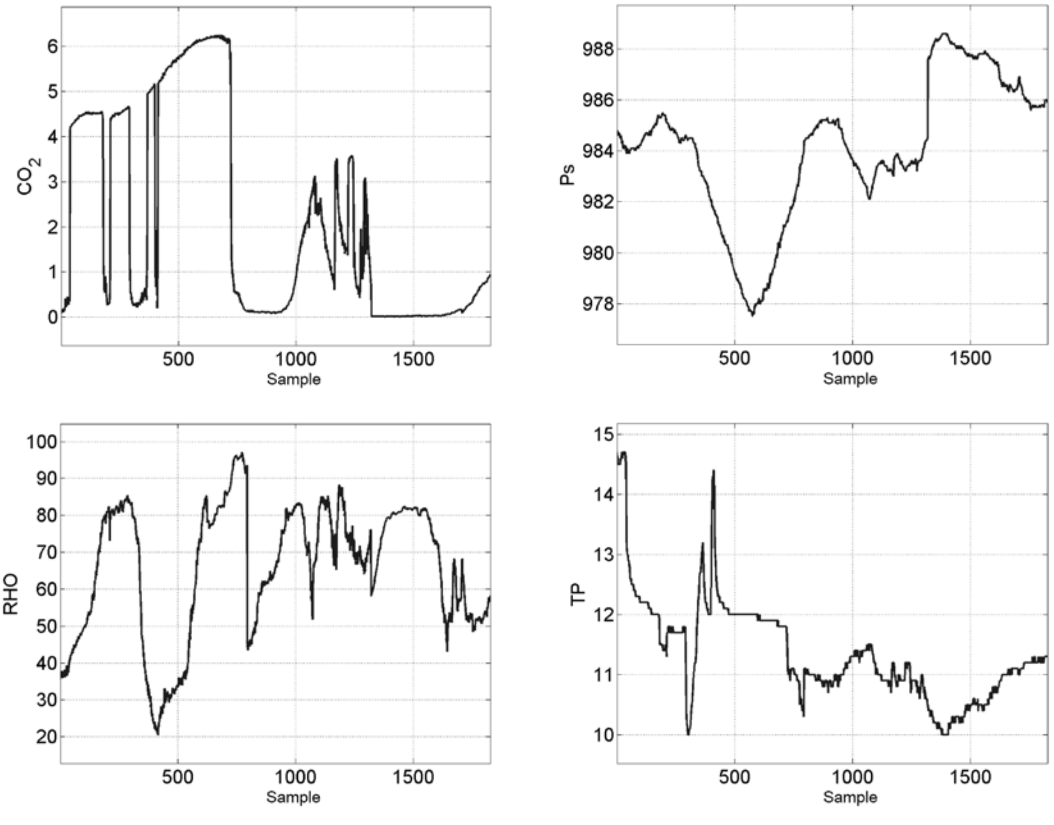


Fig. 6. Visualisation of exemplary time-series: methane concentration, air flow and mining cycle on a chosen longwall

as a parameter p) being more distant from this record then defined parameter d.

- k-NN Global Anomaly Score (GAS) – the method based on kNN approach. Each record is associated with its average distance to the rest of the records (by means of kNN method). Next, the record is identified as an outlier (or not) on the basis of interquartile range analysis.
- Local Density Cluster-Based Outlier Factor (LDCOF) – the method utilizing cluster analysis. A record is identified as an outlier on the basis of its distance to the centroid of the nearest large cluster. The distance is normalised by average distance to centroid among the members of this cluster. This method identifies small clusters as outliers.
- Histogram-based Outlier Score (HBOS) – the method based on a frequency histogram. The histogram can be created for the number of bins defined by a user or derived dynamically. The records belonging to the bin of the smaller size are labeled as outliers.

Table 5. Evaluation of outlier detection methods – random values within a given range

Algorithm	Parameters	CO$_2$ data					
		0.5%	1%	3%	0.5%	1%	3%
		Training			Testing		
GAS	0.5%	-	81.51	71.25	75.93	81.39	71.25
	1%	78.49	-	83.58	78.21	78.30	78.31
	3%	63.77	78.81	-	76.04	74.79	72.80
HBOS	0.5%	-	85.40	84.93	80.49	80.96	77.45
	1%	85.23	-	82.44	84.18	84.65	81.41
	3%	81.40	85.95	-	81.40	81.13	74.87
LDCOF	0.5%	-	81.51	76.11	75.93	81.39	71.25
	1%	78.49	-	83.58	78.21	78.30	78.31
	3%	63.77	78.81	-	76.04	74.79	72.80
CDODe	**0.5%**	-	**86.19**	**86.39**	**83.71**	**83.55**	**80.58**
	1%	**77.69**	-	**81.14**	**83.16**	**80.52**	**75.61**
	3%	**81.46**	**81.39**	-	**81.48**	**81.53**	**80.34**

A more detailed description of the methods presented above can be found in RapidMiner documentation [22].

The methods listed above were applied to the analysis of time series of measurements registered on the operator platform in mine dewatering station [24]. Each record is characterised by the following variables (see Fig. 7): CO2 – CO2 concentration on the operator platform, Ps – atmospheric pressure, RHO – humidity on the operator platform, TP – temperature on the operator platform.

In order to verify the efficiency of outlier detection methods the outlier values in quantity 0.5%, 1%, 3% of original datasets were introduced to them. The outlier values were generated with use of noise with normal distribution.

The datasets were divided into training (2/3 of original time series – initial part) and test (1/3 of original time series – last part) datasets. The task was defined as a classification one, where two classes were defined – outlier values and correct values. Due to the imbalanced distribution of the examples from the two classes the results are presented as balanced accuracy reflecting average classification accuracy in each of the classes. The value 50 means that all the examples were classified to one class what makes the method useless.

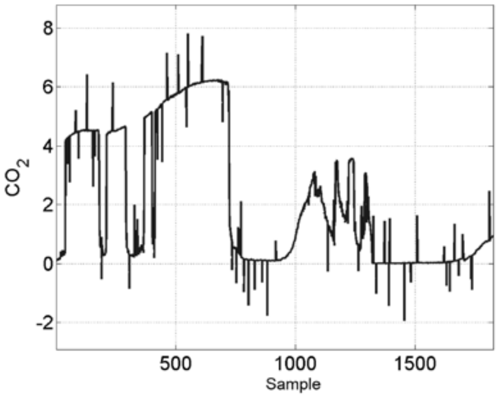


Fig. 8. CO$_2$ time series, with generated outlier values

Table 4. Evaluation of outlier detection methods – noise with normal distribution

Algorithm	Parameters	Balanced accuracy					
		0.5%	1%	3%	0.5%	1%	3%
		Training data			Test data		
GAS	0.5%	-	94.44	65.74	100	100	71.30
	1%	100	-	77.78	100	100	90.74
	3%	100	100	-	100	100	100
HBOS	0.5%	-	99.50	99.58	99.53	99.50	98.84
	1%	98.29	-	97.47	98.49	100	98.59
	3%	100	97.22	-	100	80.56	100
LDCOF	0.5%	-	100	100	100	100	100
	1%	100	-	100	99.94	99.94	100
	3%	95.07	95.24	-	95.05	94.99	96.56
CDODe	**0.5%**	-	**100**	**100**	**99.28**	**100**	**100**
	1%	**100**	-	**100**	**99.28**	**100**	**100**
	3%	**100**	**100**	-	**100**	**100**	**100**

During the first phase, where training data were analysed, the optimal parameters of the outlier detection algorithms were searched. The parameters were searched for each of the three experiments (0.5%, 1%, 3%). When the parameters were calculated, they were applied to test data analysis. The results of the analysis are presented in Table 4.

The second experiment was designed in such way that the outlier values were generated randomly from a given range encompassing the original measurements. Fig. 8 presents CO$_2$ time series containing 3% of outlier values.

The results of the analysis are presented in Table 5. It is clear that this task is much more difficult than the previous one. It can be noticed that the CDODe method is the most stable approach and it gives the best results of outlier identification.

The CDODe algorithm is a default approach to outlier identification in multidimensional time series in the DISESOR system.

5. Prediction module

Prediction module is based on, so called, prediction services. Prediction service is a webservice that predicts values of a variable (discreet or continuous) on the basis of input vector. Prediction service is inseparably connected with a model (regression or classification one) that is the basis of the prediction.

The basic scenario of prediction service application is as follows:

1. Client sends a prediction execution request accompanied by a vector of conditional attributes and a timestamp.
2. Service calculates the prediction delivering a vector of conditional attributes as a model input. The attribute values come directly from a monitoring system, because the data warehouse is not loaded online. The values of the attributes are transformed according to the dedicated ETL2 process to the form acceptable by the prediction model.
3. Service loads the results to a database.

The architecture of the Prediction module is presented in Fig. 9.

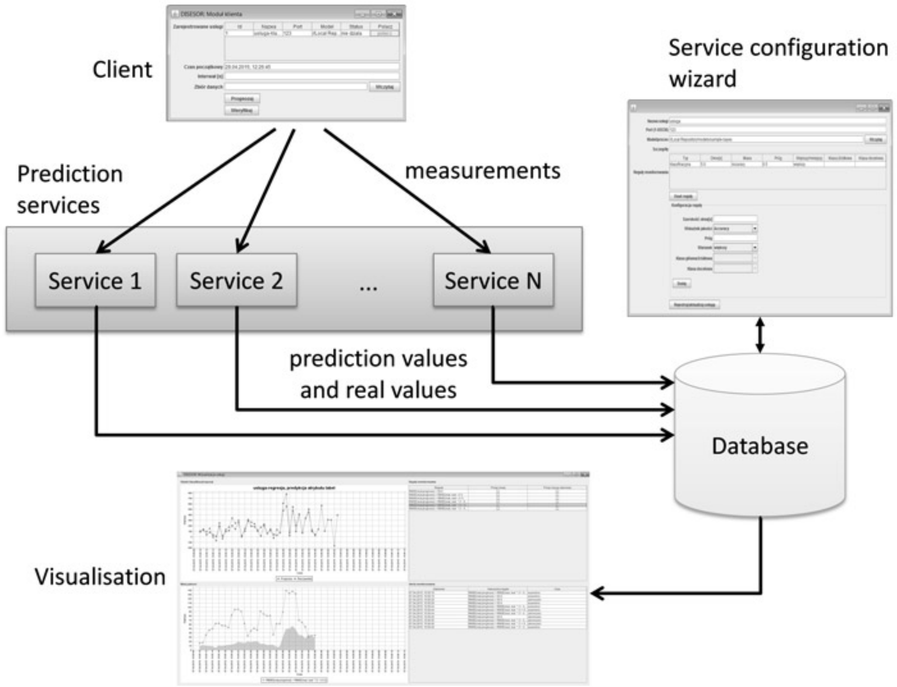


Fig. 9. Architecture and operation of prediction module

Database, which is an internal RapidMiner repository, stores the description of a model and the transformations of the attributes. Additionally, it stores the information about training data, the parameters of the minimal model quality and both predicted and real values of dependent variable. Each model adaptation results in a new database entry what makes the history of the changes available to the users.

Predictions can be visualised and compared on a single chart with the real values that are measured. Such visualisation can be performed by a monitoring or dispatching system (e.g. THOR dispatching system), where predicted values are delivered as measurements of a virtual sensor and the values of both sensors (virtual and real) can be easily compared.

It is assumed for the current module version, that if the quality of the predictions decreases below a given threshold, then a new training set is automatically collected. The size of this new data set is the same as size of the original data. The model adaptation is performed by modifying only the parameters of the existing model (the method and algorithm is not changed). Next, the quality of the model is verified on the same data that triggered the model adaptation (these data are not the part of the new training data set). If the quality of the adapted model is satisfactory, then this new model is applied to prediction. Otherwise, a message is generated stating that prediction cannot be continued and it is needed to come back to analytical module in order to create a new prediction model.

The configuration wizard enables to define the so-called quality monitoring rules. From the practical point of view there is no point in presenting the minimum model quality by means of the measures that are well-known by machine learning community, such as overall classification accuracy, g-mean, specificity, sensitivity, RMSE, MAE etc. Therefore, quality monitoring rules are based on: a sliding time-window (e.g. 1 hour) in which the quality is verified, frequency of the prediction calculation (e.g. 1 minute) and the indicators which are typically called *FalsePositive* and *FalseNegative*. The values of these indicators are explicitly defined by a user for each decision class or only for a target class, e.g. corresponding to “danger”. Therefore, knowing the values of *FalsePositive* and *FalseNegative* [3], and a number of predictions that are calculated in a given time-window it is possible to calculate the values of almost all the possible quality measures of prediction model. In case of regression task the module allows so-called insensitivity, what means that the predictions that differ less than the given threshold from the real values are not treated as an error. Additionally, it is possible to define that the values within the given range (e.g. corresponding to the “normal” state) are not counted as errors.

6. Example of the system application to the task of abyssal mining pump stations diagnostics

Abyssal mining pump stations represent a fundamental solution to the problem of a coal mine dewatering. Due to the large responsibility in maintaining the water at a certain level, that guarantees the safe operation of the mine, the systems that oversee the abyssal mining pump stations are safety systems. The pump monitoring systems are installed in several dewatering stations and during the normal operation they register the following pump unit parameters:

- pump unit temperature,
- the power consumed by the motor,
- the current drawn by the motor,
- the productivity of a pump unit.

The values of the parameters listed above are acquired each second. Due to the safety constraints the temperature of the pump motor should not exceed 75 °C and a pump should be turned on when its temperature decreases below 25 °C.

Each underground water well contains four pumping units (see Fig. 10).

Analysis of the collected measurements enables the evaluation of the pump diagnostic states. The following feature vector was used during the analysis of pump diagnostic states:

$$P_i = [T_{U,i}\,,\ T_{0,i}\, t_{20\text{-}30,i},\ t_{30\text{-}40,i},\ t_{40\text{-}50,i},\ t_{50\text{-}60,i},\ t_{60\text{-}70,i},\ P_{U,i}\, Q_{U,i},\ L_i,\ D_{pi},\ D_{ki}],$$

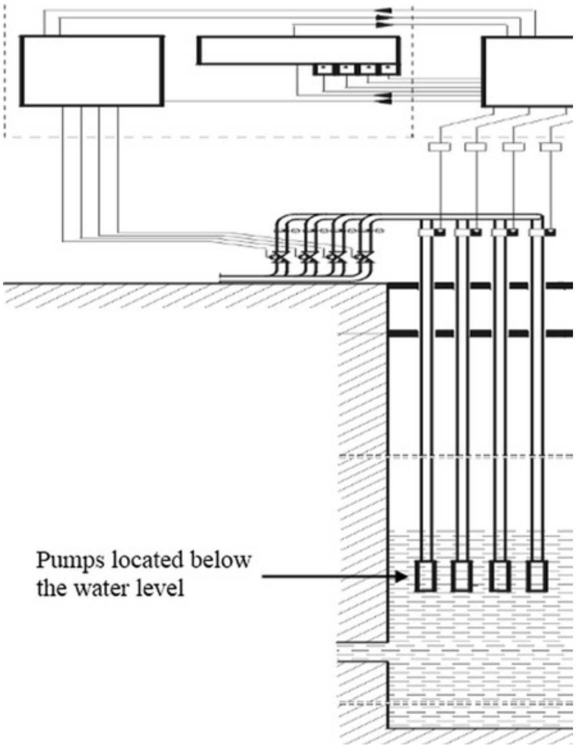


Fig. 10. Abyssal mining pump station

where:

- T_U – temperature of a pump unit in a steady state,
- T_0 – initial temperature of a pump unit,
- t_{x-y} – time period when the pump temperature changes by 10 degrees (if the pump temperature has not reached a given range, a 0 value was inserted),
- P_U – power of a pump unit in a steady state,
- Q_U – performance of a pump unit in a steady state,
- L – number of starts on the previous day,
- D_p – time and date when the pump was turned on,
- D_k – time and date when the pump was turned off.

Temperature of a pump unit in a steady state (T_U) was calculated as an average value of the last two minutes of operation. Each record P_U reflects a single pumping cycle (starting from the unit turn on to turn off).

Analysis of historical data and interview with the dispatchers of the station (experts) enabled to define three diagnostic states: *a new pump unit* (also after repair), *correct operation* and *suitable for repair*.

The main impact on the diagnostic state of a pump unit have time periods t_{x-y}. Along the pump unit operation, when it becomes exploited, the time periods t_{x-y} become shorter and the critical temperature when the pump must be turned off is reached faster. This results in pump numerous turn on during the days preceding decision of its repair. Therefore, the number of times the pump was turned on is an important diagnostic indication. It has to be regarded, however, in conjunction with the information about the temperature of a pump unit in a steady state in order to omit other than high temperature turn off reasons.

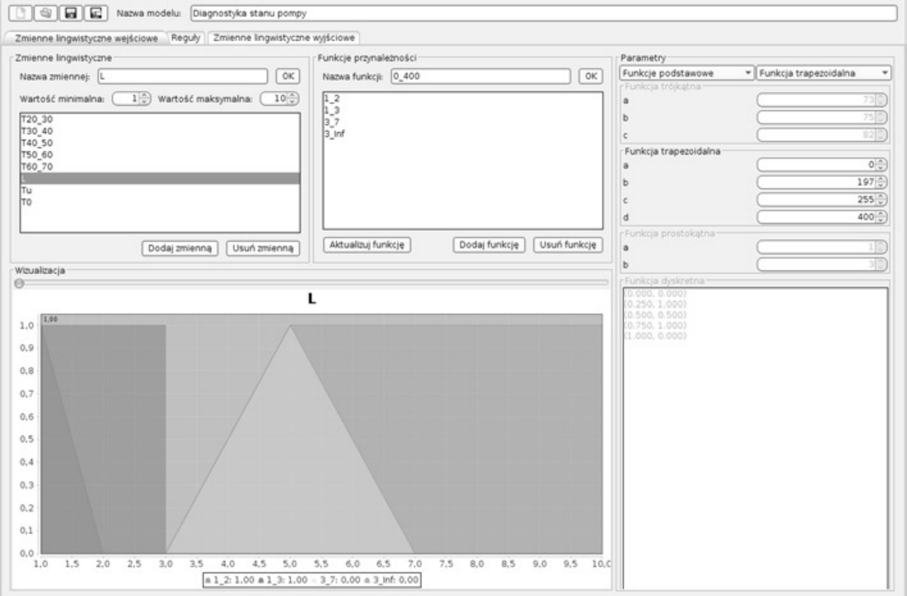
Pump state diagnostics was based on a Mamdani-type fuzzy system [16] with the following rules (the notation (p1, p2, p3, p4) reflects trapezoidal membership function):

IF T20_30 ∈ (199, 255, 255, 409) THEN a new pump unit
IF T20_30 ∈ (0, 197, 255, 409) and
T30_40 ∈ (245, 246, 256, 362) and
T60_70 ∈ (826, 1159, 1473, 679715) THEN a new pump unit

IF T20_30 = 0 and L∈(1, 1, 1, 2) THEN correct operation
IF T20_30 = 0 and
T40_50 ∈ (0, 0, 387, 727) and
L ∈ (1, 1, 3, 3) THEN correct operation
IF T20_30 ∈ (0, 255, 255, 409) and
T_u ∈ (73.1, 73.41, 74.54, 81.9) THEN correct operation

IF T20_30 = 0 and
T50_60 ∈ (0, 390, 390, 551) and
L ∈ (3, 5, 5, 7) THEN suitable for repair
IF T_0 ∈ (15.14, 19.75, 19.75, 27.26) and
T20_30 = 0 and
T30_40 ∈ (206, 366, 366, 11417) and
L ∈ (3, 5, 7, 7) THEN suitable for repair

Fig. 11 presents the division of attribute L (number of starts on the previous day) into fuzzy sets.


Nazwa modelu: Diagnostyka stanu pompy

Zmienne lingwistyczne wejściowe | Reguły | Zmienne lingwistyczne wyjściowe

Zmienne lingwistyczne
Nazwa zmiennej: L
Wartość minimalna: 1 Wartość maksymalna: 10
T20_30
T30_40
T40_50
T50_60
T60_70
L
Tu
T0

Funkcje przynależności
Nazwa funkcji: 0_400
1_2
1_3
3_7
3_Inf

Parametry
Funkcje podstawowe
Funkcja trapezoidalna
Funkcja trójkątna
a 73
b 75
c 82
Funkcja trapezoidalna
a 0
b 197
c 255
d 400
Funkcja prostokątna
a 1
b 3
Funkcja dyskretna
(0.000, 0.000)
(0.250, 1.000)
(0.500, 0.500)
(0.750, 1.000)
(1.000, 0.000)

Wizualizacja
1_2: 1.00 1_3: 1.00 3_7: 0.00 3_Inf: 0.00

Fig. 11. presents the division of attribute L (number of starts on the previous day) into fuzzy sets.

In practice, the state *suitable for repair* does not lead to immediate brake down of a pump unit. Dispatchers suggested that a pump classified to this state is able to (or sometimes has to – waiting for a service) operate up to next 3 months (when a typical operation time lasts 2 years). In order to improve the accuracy of service prediction (pump break down) a decision tree was created by means of a decision tree induction algorithm. Tree induction was performed only on the examples labeled as *suitable for repair*. The resulting tree classifies each vector P_i to one of two decision classes: *less than a month to break down* and *more than a month to break down*. The induced tree utilises only the time periods t_{x-y}. An applied train-and-test method showed the classification accuracy on a level of 90% (the class distribution was balanced, therefore, this measure is appropriate to clas-

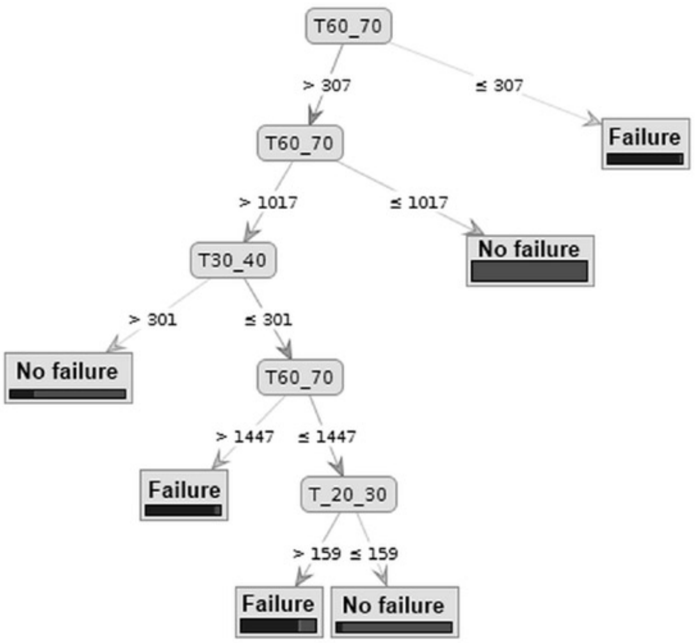


Fig. 12. The decision tree applied to the expert system

sification quality evaluation). The induced tree (the main node) was slightly modified, what increased accuracy by 2%, and it was applied to the expert system. The decision tree that was induced is presented in Fig. 12, where decision classes: *less than a month to break down* and *more than a month to break down* are represented as *Failure* and *No failure* respectively.

The expert system works according to the following steps: after each pumping operation the diagnostic state of a pump unit is evaluated. If a pump is classified as *suitable for repair*, then the decision tree is applied. It identifies the expected time period of the pump operation. If the pump is classified as *less than a month to break down* then the pump should be turned off because the costs of the repair of the broken unit are very high.

7. Example of the system application to the task of methane concentration prediction in mining excavation

The DISESOR system can be applied to solve a variety of tasks. This section presents an example, of the system application to methane concentration prediction.

Methane concentration monitoring is one of the main tasks of the natural hazard monitoring systems in mining industry. Such system is in charge of automatic and immediate shut-down of electricity within a given area, if a methane concentration exceeds a given alarm threshold. The power turn-on is possible after a certain time (from 15 minutes to even several hours), when the methane concentration decreases to the acceptable level. This results in large losses associated with downtime of production. Information from a soft (virtual) sensor presenting to a dispatcher the prediction of the methane concentration with a few minute horizon can prevent electricity shut-down or can allow to lower the mining activity and increase the air flow if possible. Therefore, these actions allow to avoid undesirable situations and unnecessary downtimes.

The task of maximal methane concentration prediction with the horizon from 3 to 6 minutes was realised within the DISESOR system. By means of ETL2 module a set of the following sensors was selected: AN321, AN541, AN547, AN682, BA1000, BA603, BA613, BA623, MM11, MM21, MM25, MM31, MM36, MM38, MM39, MM41, MM45, MM52, MM53, MM54, MM55, MM57, MM58, MM59, MM61, MM81.

The data were aggregated applying minimum operation to anemometer (AN) measurements, average operation to barometer (BA) measurements and maximum operation to methanometer (MM) measurements. The missing values were imputed applying linear regression method. As a dependent variable MM59 sensor was chosen. A map presenting the topology of the mining area and location of the sensors is presented in Fig. 13.

As analytical module is currently being developed, the method of regression tree induction was chosen arbitrarily to create the prediction model. The initial tree was created on the basis of data coming from 1 shift. The model and the list of sensors (variables) together with the defined transformations were forwarded to prediction model running a proper service. The time-window defined for prediction quality monitoring was set to 1 hour and the model adaptation was executed each hour regardless the minimum quality requirements. The adaptation could be executed more often if the minimum quality requirements were not met but there was no such situation. The data that were predicted were delivered on-line by the simulator of THOR system in order to simulate the real stream of measurements.

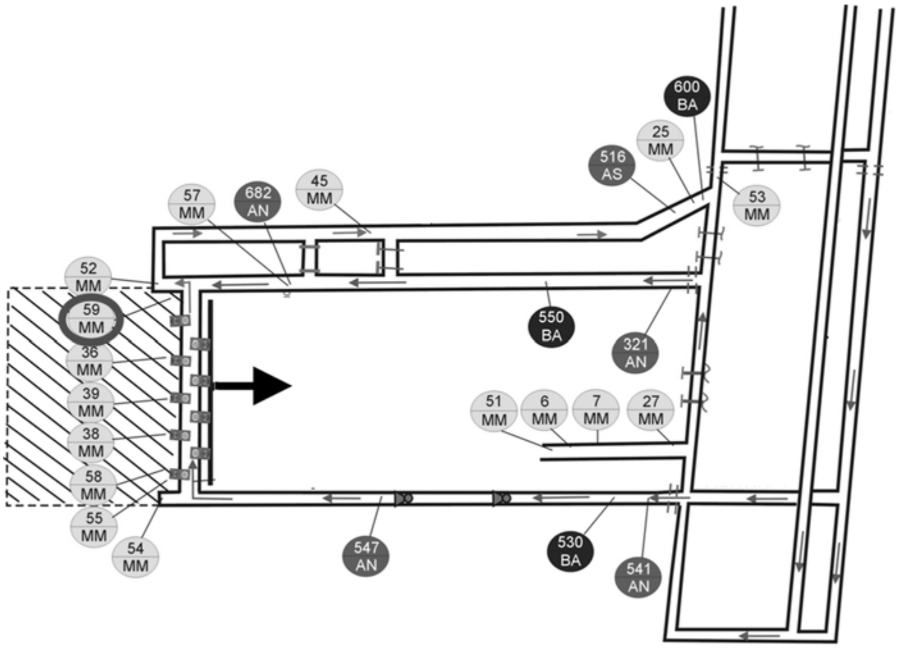


Fig. 13. Topology of the mining area and location of the sensors – MM59 sensor chosen as dependent variable is outlined a thick line

Fig. 14 presents the process of data preparation and the prediction model creation together with the initial regression tree that was created. Whereas, Fig. 15 presents the plot of the real methane concentration and the predicted maximum concentration together with the histogram of errors that are reported to a user. Currently, the user interface is in Polish as the system deployment in Poland was planned in the project. However, the English and Chinese versions are also planned.

Fig. 15. The plot of the real methane concentration and the predicted maximum concentration together with the histogram of errors that are reported to a user

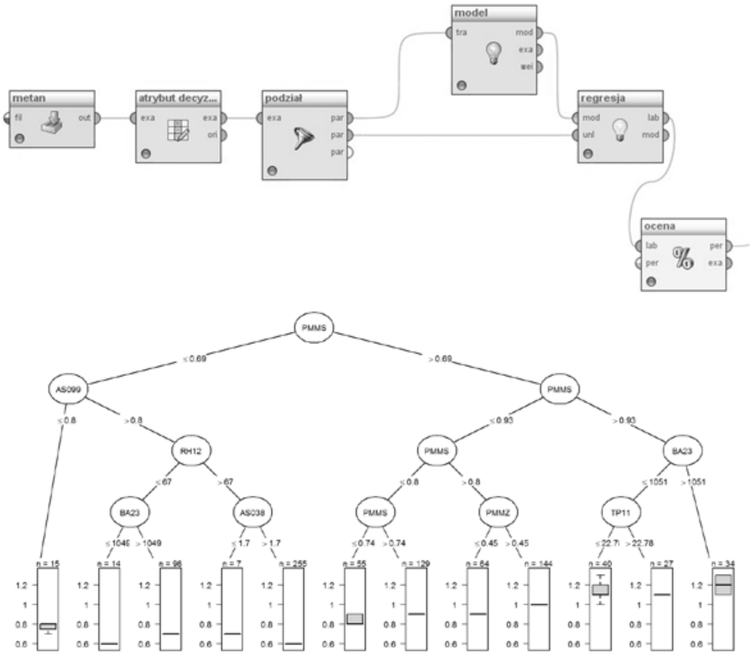


Fig. 14. The process of data preparation and prediction model creation together with the initial regression tree that was created

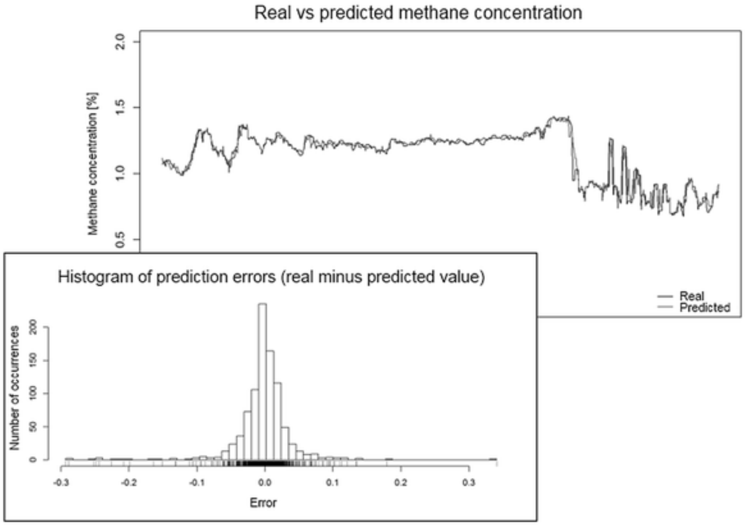


Fig. 15. The plot of the real methane concentration and the predicted maximum concentration together with the histogram of errors that are reported to a user

8. Conclusion

The system that is being developed delivers the solutions for decision support of a dispatcher and process operator. This system is complete as it delivers the tools that can be applied to data storage, processing and preparation, and also to definition of the models based on expert knowledge (expert system) and the models based on the results of both historical and on-line data analysis. Due to the application and proper customisation of existing tools (RapidMiner, R) and development of the proprietary solutions (e.g. ETL2, rule induction and rough set operators [23] that are not available in RapidMiner) a user receives a broad set of tools that can be applied to different tasks. Finally, the case studies that were presented show that the system can be practically utilised in a coal mine industry.

The DISESOR system provides analytical tools available for advanced users as well as for users who are not data analysts (through many wizards that facilitate the use of the system). However, a routine use of the system requires, in our opinion, a new, gaining popularity, workplace, which is data scientist [2].

Acknowledgement

This research was supported by Polish National Centre for Research and Development (NCBiR) grant PBS2/B9/20/2013 in frame of Applied Research Programme. The work was also financially supported by POIG.02.03.01-24-099/13 grant: GeCONiI Upper Silesian Center for Computational Science and Engineering.

References

1. Bifet A, Holmes G, Kirkby R, Pfahringer B. Moa: Massive online analysis. The Journal of Machine Learning Research 2010; 11: pp. 1601-1604.
2. Dhar V. Data science and prediction. Communications of the ACM 2013; 56(12): 64-73, http://dx.doi.org/10.1145/2500499.
3. Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006; 27(8): 861-874, http://dx.doi.org/10.1016/j.patrec.2005.10.010.
4. Gąsior S. Diagnosis of longwall chain conveyor. Przegląd Górniczy 2001; 57(7-8): 33-36
5. Głowacz A. Diagnostics of Synchronous Motor Based on Analysis of Acoustic Signals with the use of Line Spectral Frequencies and K-nearest Neighbor Classifier. Archives of Acoustic 2015; 39(2): 189-194, http://dx.doi.org/10.2478/aoa-2014-0022.
6. Grzegorowski M. Scaling of complex calculations over big data-sets. Active Media Technology, Springer 2014; pp. 73-84.
7. Jurdziak L, Zimroz R. Dlaczego diagnostyka maszyn się opłaca i ile można na tym zaoszczędzić? Prace Naukowe Instytutu Górnictwa Politechniki Wrocławskiej 2004; 106: 139-150.

8. Kabiesz J. Effect of the form of data on the quality of mine tremors hazard forecasting using neural networks. Geotechnical & Geological Engineering 2006; 24(5): 1131-1147, http://dx.doi.org/10.1007/s10706-005-1136-8.
9. Kabiesz J, Sikora B, Sikora M, Wróbel Ł. Application of rule-based models for seismic hazard prediction in coal mines. Acta Montanistica Slovaca 2013; 18(4): 262-277.
10. Kacprzak M, Kulinowski P, Wędrychowicz D. Computerized information system used for management of mining belt conveyors operation. Eksploatacja i Niezawodność - Maintenance and Reliability 2011; 2(50): 81-93.
11. Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. Computers & Chemical Engineering 2009; 33(4): 795-814, http://dx.doi.org/10.1016/j.compchemeng.2008.12.012.
12. Kalisch M, Przystalka P, Timofiejczuk A. Application of selected classification schemes for fault diagnosis of actuator systems. Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on. IEEE 2014; 1381-1390, http://dx.doi.org/10.15439/2014f158.
13. Kimball R, Ross M. The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons, 2011.
14. Kozan E, Liu S Q. A demand-responsive decision support system for coal transportation. Decision Support Systems 2012; 54(1): 665-680, http://dx.doi.org/10.1016/j.dss.2012.08.012.
15. Leśniak A, Isakow Z. Space-time clustering of seismic events and hazard assessment in the Zabrze-Bielszowice coal mine, Poland. International Journal of Rock Mechanics and Mining Sciences 2009; 46(5): 918-928, http://dx.doi.org/10.1016/j.ijrmms.2008.12.003.
16. amdani E H, Assilian S. An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies 1975; 7(1): 1-13, http://dx.doi.org/10.1016/S0020-7373(75)80002-2.
17. March S T, Hevner A R. Integrated decision support systems: A data warehousing perspective. Decision Support Systems 2007; 43(3): 1031-1043, http://dx.doi.org/10.1016/j.dss.2005.05.029.
18. Mazurkiewicz D. Computer-aided maintenance and reliability management systems for conveyor belts. Eksploatacja i Niezawodnosc - Maintenance and Reliability 2014; 16(3): 377-382.
19. Michalak M, Sikora M, Sobczyk J. Analysis of the longwall conveyor chain based on a harmonic analysis. Eksploatacja i Niezawodność - Maintenance and Reliability 2013; 15(4): 332-333.
20. PostgreSQL. Postgresql 2015; Available online: http://www. postgresql.org/.
21. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria 2014; Available online: http://www.R-project.org
22. RapidMiner. RapidMiner 2015; Available online: http://rapidminer.com
23. Riza L S, Janusz A, Bergmeir C, Cornelis C, Herrera F, Ślęzak D, Benitez J M. Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets". Information Sciences 2014; 287: 68-89, http://dx.doi.org/10.1016/j.ins.2014.07.029.
24. Sevitel. Thor 2015; Available online: http://www.sevitel.pl/product,25,THOR.html
25. Sikora M, Michalak M. Eksploracja baz danych systemów monitorowania na przykładzie obserwacji pracy kombajnu chodnikowego. Bazy Danych: Rozwój metod i technologii, (Tom 1: Architektura, metody formalne i zaawansowana analiza danych), WKŁ, Warszawa 2008; 429-437.
26. Sikora M, Sikora B. Improving prediction models applied in systems monitoring natural hazards and machinery. International Journal of Applied Mathematics and Computer Science 2012; 22(2): 477-491, http://dx.doi.org/10.2478/v10006-012-0036-3.
27. Sikora M, Sikora B. Rough natural hazards monitoring. Rough Sets: selected Methods and Applications in Management and Engineering, Springer 2012; 163-179, http://dx.doi.org/10.1007/978-1-4471-2760-4_10.
28. Simiński K. Rough subspace neuro-fuzzy system. Fuzzy Sets and Systems 2015; 269: 30-46, http://dx.doi.org/10.1016/j.fss.2014.07.003.
29. Talend. Talend open studio 2015 Available online: https://www.talend.com/ products/talend-open-studio.
30. Wachla D, Moczulski W A. Identification of dynamic diagnostic models with the use of methodology of knowledge discovery in databases. Engineering Applications of Artificial Intelligence 2007; 20(5): 699-707, http://dx.doi.org/10.1016/j.engappai.2006.11.002.
31. Zimroz R. Metody adaptacyjne w diagnostyce układów napędowych maszyn górniczych. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2010.

Michał KOZIELSKI
Institute of Electronics
Silesian University of Technology
Akademicka 16, 44-100 Gliwice, Poland

Marek SIKORA
Institute of Informatics
Silesian University of Technology
Akademicka 16, 44-100 Gliwice, Poland

Łukasz WRÓBEL
Institute of Innovative Technologies EMAG
Leopolda 31, 40-189 Katowice, Poland
Institute of Informatics, Silesian University of Technology
ul. Akademicka 16, 44-100 Gliwice, Poland

E-mails: michal.kozielski@polsl.pl, marek.sikora@polsl.pl, lukasz.wrobel@ibemag.pl