



# Analytics over Multi-sensor Time Series Data – A Case-Study on Prediction of Mining Hazards

Andrzej Janusz<sup>1,2</sup>(✉)  and Dominik Ślęzak<sup>1,2</sup> 

<sup>1</sup> Institute of Informatics, University of Warsaw, Warsaw, Poland

{janusza, slęzak}@mimuw.edu.pl

<sup>2</sup> QED Software, Warsaw, Poland

**Abstract.** Mining of high-dimensional time series data that represent readings of multiple sensors is a challenging task. We focus on several important aspects of analytics over such data. We describe a methodology for extracting informative features from multidimensional data streams, as well as algorithms for finding compact representations of such data, in order to facilitate the construction of prediction models. We pay special attention to designing new approaches to dimensionality reduction and interchangeability of features that such representations comprise of. We validate our algorithms on data sets obtained from coal mines and we demonstrate how their results can be applied for a construction of a decision support system. We show that such system is efficient and that its outcomes can be easily interpreted by subject matter experts.

**Keywords:** Decision support systems · Time series data · Data representation learning · Dimensionality reduction · Interchangeability of features

## 1 Introduction

Multiple-sensor data streams have become one of the main sources of big data. Their applications can be found in security systems, health care, scientific data mining or monitoring of industrial processes, etc. Analysis of such data is a challenging task that often requires integration of specialized hardware with efficient data processing algorithms. This is particularly important for systems aimed at monitoring safety conditions and hazards in underground mines.

The problem of selecting an appropriate data representation for analysis is fundamental in the most of typical KDD processes. Particularly, in the domain of multidimensional time series data, this task can be approached from many different angles. On the one hand, a desirable data representation at every given point of time should comprehensively describe a state of the observed environment or phenomenon. On the other hand, the representation should be concise

to facilitate data processing and construction of analytic models. For those reasons, one of the most common approaches divides this task into two steps. Firstly, a large number of features is defined, whose values can be computed for each point of time. Such features usually correspond to some descriptive statistics and aggregations obtained from values of the time series, recorded in a period of time that directly precedes the given moment. Since the features constructed in the first step are often redundant, in the second step, feature selection methods are applied to reduce the dimensionality of the final representation.

In applications related to predictive analysis of mining hazards, it is also substantial to ensure fail-safety measures and interpretability of resulting models. This can be achieved by working on small, yet informative feature sets and using relatively simple prediction models. To improve robustness of the results, it is also appealing to identify which features and data sources provide similar information and can be regarded as interchangeable. In this paper, we demonstrate a framework for constructing reliable and simple predictive models. We discuss two real-life case studies corresponding to utilization of various sensors placed in underground coal mines to monitor potential hazards to miners and their equipment. We particularly analyze importance and interchangeability of features extracted from multidimensional time series data. We also compare the performance of the constructed models with the state-of-the-art.

The paper is divided into six sections. In Sect. 2, we discuss two examples of challenging problems of the coal mining industry, namely predicting dangerous concentrations of methane and early detection of periods of increased seismic activity. In Sect. 3, we consider the task of extracting informative data representation from streams of sensor readings. We describe features that we use in our models and propose a procedure for feature ranking that is inspired by the rough set theory. In Sect. 4, we propose a heuristic for detecting interchangeable features and explain how it can be used for a purpose of constructing interpretable models. Section 5 presents results of experiments in which we tested the described feature extraction and predictive modeling chain on the data from two international data mining competitions. Section 6 concludes the paper.

## 2 Related Work

Coal mines are usually well equipped with monitoring and dispatching systems that are connected with all machines, devices and transport facilities used underground. Specialized systems exist also for monitoring risks of natural hazards, such as methane explosions and seismic tremors [3]. These systems are usually provided by different companies, which causes problems related to the quality, integration and correct interpretation of the data that they collect. The collected data sets are used primarily for creating temporary (current) visualizations on boards that display certain locations in the mine. Such boards require constant attention of humans who supervise the coal mining process and assess related hazards. Moreover, the commonly used systems have a very limited set of analytical tools available to subject matter experts, whereas an appropriate

combination of domain knowledge with results of historical data analysis could considerably facilitate various kinds of mining processes [11].

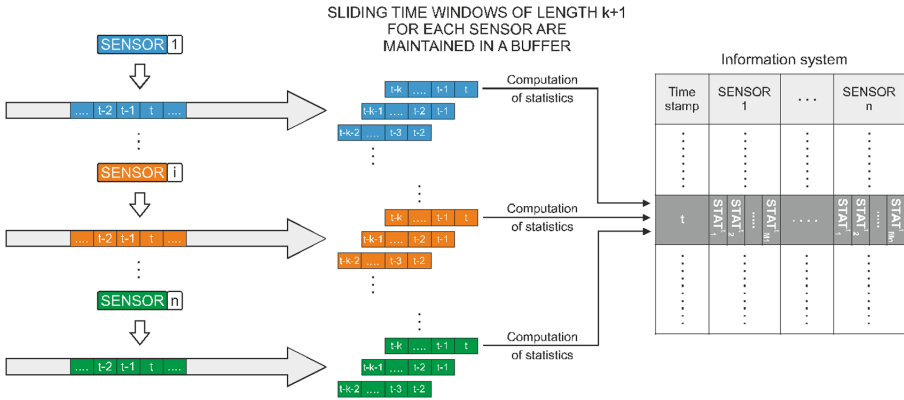
In this paper, we give examples of challenging real-life problems related to the mining industry and propose an approach to solving them using techniques that are efficient and scalable, and on the other hand are comprehensible for subject matter experts. In practice, such methods are especially important due to the fact that their results can be easily interpreted and validated. The first of our case studies is related to the problem of predicting dangerous concentrations of methane in a longwall of a coal mine. In practice, this is a serious concern, as methane gas is highly explosive at a concentration of 4.4%–15% and has been responsible for many deadly mining disasters<sup>4</sup>. Another fundamental problem that occurs in coal mines is related to seismic tremors and the early detection of periods of increased seismic activity [1,9]. Both of those tasks were in a scope of data mining competitions held on the Knowledge Pit platform (<https://knowledgepit.ml>).

In the first competition, *Mining Data from Coal Mines*, participants were asked to construct a models for predicting warning levels of methane concentrations on three methanometers located along a longwall of an active mine. The available data consisted of series of readings (10 min long) from 28 sensors. Apart from the concentration of methane, sensors monitored various other conditions in the mine (air flow speed, intensity of cutter loader work, etc.) [15]. The second competition, *Predicting Dangerous Seismic Events in Active Coal Mines*, aimed at evaluating the efficiency of expert methods for the assessment of seismic hazards in underground mines [6]. Herein, participants were constructing models for predicting seismic activity based of aggregated readings from seismic sensors (e.g.: geophones), additional expert knowledge related to particular mining sites (the data came from 24 different sites), and mining operations (e.g.: preventive blasts). In both cases, the available data could be characterised as multidimensional time series and thousands of features can be defined to describe each time point. However, the analysis of the most successful solutions revealed that robust predictive models for those tasks may use only a relatively small number of features, in order to improve their reliability and increase the interpretability of their outcomes. More information regarding these competitions and the used data sets can be found on their respective web pages.

### 3 Representation of Multidimensional Time Series Data

In both case studies considered in this paper the available data sets consisted of sensor readings recorded over a period of time, whereby target classes associated with each of time points indicated a presence or absence of a condition that is determined by future values of one of the monitored sensors. Those two common characteristics of the investigated tasks led us to design a single comprehensive feature engineering framework that can be used for constructing informative representation of multidimensional time series data. In our framework, at each

<sup>4</sup> <http://mining.about.com/od/Accidents/a/Common-Mining-Accidents.htm>.



**Fig. 1.** A schema showing how sensory data streams are processed in our framework.

point in time there is considered a period that directly precedes this moment. Such a period is typically called a sliding time window due to a fact that its length is fixed. Readings of all sensors recorded within a time window constitute a basis for computation of features that represent the corresponding moment in time. Figure 1 illustrates the corresponding schema.

Many different types of features can be constructed in our feature engineering framework. Basically, any typical aggregation function can be efficiently computed over values from time windows that are stored in a special buffer for every available sensor. Common examples of such functions are min, max, mean, standard deviation and quantiles. All these statistics can also be computed over non-overlapping sub-windows of the current window. For instance, if a time window stores sensor readings from the latest 10 min, the basic statistics can be computed independently for each consecutive two minutes, which gives additional five features. Moreover, new features may include information regarding latest occurrences of some special events, e.g., how long ago the min/max value was recorded in the current time window. It is also possible to investigate statistics of transformed window values, such as a max value of differences between consecutive sensor readings. Finally, newly derived features may express dependencies between readings (or precomputed statistics) of different sensors.

This way, it is possible to construct a vast amount of features that constitute an informative representation of a given moment in time. We decided to use the same types of features to represent the data for both of the considered case studies. In particular, for readings of each sensor, we used a set of features that are briefly outlined in Table 1. The processed time series data sets were stored in a form of flat data tables. Due to a diversity of extracted features and a high number of considered sensors, this data representation was high-dimensional. For the methane-related data, the total number of extracted features was 2100. For seismic data, it was 743, out of which 726 were extracted from time series and 17 corresponded to general assessments of conditions in a given moment. These

17 features included latest safety assessments provided by experts, meta-data of the corresponding working sites and the most recent measurements that are typically done only once in every 24 h (or even less frequently).

**Table 1.** A list of features that were used to represent each time window. The same feature types were used for the problem of predicting warning levels of methane concentrations and foreseeing periods of increased seismic activity. All these features were computed independently for each of monitored sensors.

Feature type	Description
Basic summary of all readings in a given time window	Simple statistics: mean, median, min, max, stdDev, 10th and 90th percentiles
Summaries of non-overlapping sub-windows	The same statistics as above, computed over sub-windows (2 mins for methane and 8 h for seismic activity data)
Statistics focused on a trend of the most recent readings in a time window	Differences between the last and min/max values, between the last and the first value, between corresponding percentiles (10th, 50th and 90th) of two latest sub-windows
Statistics computed over differentiated values in a time window	Mean, median, min, max, stdDev, 10th and 90th percentiles of differences between two consecutive readings in a time window
Statistics computed over transformed (aggregated) values of a time window or consecutive sub-windows	Differences between min and max values, mean and median values, and max and 90th/min and 10th percentiles, differences between the last two mean values for consecutive sub-windows
Indicators of extreme readings in a time window and sub-windows	Position in the time window (a relative time index) of a reading with min/max value, position of a min/max value in the latest sub-window
Indicators of extreme values computed over transformed readings	Position of a min/max difference between consecutive readings, position of min/max of means computed from consecutive sub-windows

Given such large number of constructed features, the obtained representation of time windows is likely to be redundant. Thus, a proper selection of features could be even more beneficial than utilization of sophisticated prediction methods. A desirable method of selecting a feature subset needs to meet several requirements dictated by the considered application area, namely efficiency and scalability, robustness with respect to changes and data artifacts, and independence from a particular model or a prediction approach.

The first requirement is obvious considering the fact that in both of the investigated problems recorded data streams may involve multiple different sensors, and a frequency of readings can also be high. In a general case, the size of a set of extracted features can be unlimited, thus the resulting representation of each time window can be extremely high-dimensional. The second requirement is motivated by constantly changing conditions in underground coal mines. For instance, a working longwall shearer changes its position while making progress in cutting the coal from a coalface. For this reason, some of sensors can be relocated, which in turn may alter properties of their readings. Additionally, severe conditions during the mining process often cause some hardware malfunctions and enforces regular maintenance and calibration of sensory equipment. The third requirement is dictated by the fact that it is much more convenient to indicate a subset of features that are meaningful regardless of a method applied to construct a particular prediction model. Not only does it give a flexibility to work with and compare different analytical methods but it also makes the final models easier to interpret by subject matter experts.

From many possible approaches to feature selection, the most suitable in our case seems to be the one that is based on multivariate filtering techniques. This family consists of algorithms that create rankings of attributes and use them to identify a subset that is likely to contain only relevant features. In order to achieve that, each of features is given a score that expresses its usefulness in a context of other features. Quite often, to select a compact subset of meaningful features that represent different aspects of considered data objects, multivariate attribute rankings are combined with heuristics for finding diverse features, e.g., the minimal redundancy-maximal relevance approach (*mRMR*) [13].

Another prominent example of a family of multivariate filtering methods was developed within the rough set theory, whereby the feature selection task typically corresponds to the notion of a reduct. It can be defined as a minimal subset of attributes that preserves information about objects with different properties, e.g., having different target labels [12]. Heuristics for computing different types of reducts have been widely discussed. There were also proposed several extensions to reducts that aimed at dealing with a problem of illusionary dependencies<sup>2</sup>. One of such extensions is the notion of approximate reducts [16]. It allows a small loss of information regarding target labels in exchange for better tolerance of noisy or randomly disturbed data. It is worth to notice that all reduct computation algorithms are in fact using the multivariate filtering approach, since their attribute quality criteria always assess a given feature in a context of information provided by a set of already chosen attributes.

Out of many known heuristics for computation of approximate reducts, we follow so-called DAAR algorithm that is based on iterative selection of the most useful feature in a context of previously selected attributes [7]. At each step, a quality measure (e.g.: entropy gain) is used to assess a random subset of features and the one with the highest score is taken to the resulting subset. In DAAR, the

---

<sup>2</sup> By illusionary we mean dependencies that occur in a given data sample but they would not occur if more complete information was provided.

stopping criterion uses a permutation test to automatically adjust the approximation threshold of constructed reduct, in order to avoid selecting attributes that are likely to discern objects only by chance. The algorithm uses a Monte Carlo procedure during the selection of an optimal feature at each iteration, thus it can be used to find many different approximate reducts for the same input data. An implementation of this heuristic and a few other reduct computation methods are available in the R System library *RoughSets* [14]<sup>3</sup>.

Reduct computation methods can also be used to define a ranking of features with regard to their usefulness in discriminating data cases with different target labels. The most straightforward of reduct-based ranking algorithms constructs such rankings on a basis of a frequency with which a feature appears in a reduct computed using the Monte Carlo method [10]. A feature that is often included in a reduct must have been assessed as useful in many contexts defined by different subsets of other features, and thus is likely to be truly relevant. In Sect. 4, we demonstrate that even this simple method, combined with DAAR and detailed analysis of pairwise feature interchangeability, can yield meaningful feature rankings and facilitate construction of reliable prediction models.

## 4 Interchangeability of Features in Hazard Assessment

Usually, two features are regarded interchangeable if information that they provide is sufficiently similar, so that if values of one feature are known, then the other one can be safely assumed as redundant. Information about such features is valuable in practice. For instance, it can be directly utilized for a purpose of feature selection process. Moreover, it can be utilized to increase robustness of predictions issued by a decision support system due to a possibility of constructing concurrent prediction models that use different, yet pairwise interchangeable feature sets. Not only can it boost the prediction quality (by using an ensemble) but it also increases reliability of the system in case when one of sensors generating input data stopped working properly.

There exist a number of methods for measuring a similarity of information conveyed by a pair of features [8]. The most common is based on linear correlation, i.e., the higher is the absolute value of correlation, the more similar two attributes are presumed to be. However, this approach can handle only monotonic dependencies and can be applied only to numerical data. We propose a different way of detecting interchangeable features that is based on a notion of co-occurrence in approximate reducts. Intuitively, it is expected that two features are similar if they rarely appear in the same reduct and they often co-occur with the same subsets of other features. If two feature are often in the same reduct it means that they must distinguish different subsets of data cases, and thus information that they express is different. However, if two features often appear in different reducts with similar subsets of other attributes, it means that they add comparable information in many different contexts.

---

<sup>3</sup> <https://github.com/janusza/RoughSets>.

To formalize this observation, let us consider a finite set of approximate reducts  $R$  computed for a data table  $\mathbb{S}_d = (U, A \cup \{d\})$ , where  $U$  is a set of data cases,  $A$  is a set of their features and  $d$  is the target variable. Additionally, by  $AR_k$  let us denote  $k$ -th element of the set  $R$ . For any  $a_i, a_j \in A$ , we may define a frequency of their co-occurrence in reducts from  $|R|$ , i.e.:

$$f_{i,j} = \frac{|\{k : a_i \in AR_k \wedge a_j \in AR_k\}|}{|\{k : a_i \in AR_k\}|} \tag{1}$$

Values of  $f_{i,j}$  can be arranged into a matrix  $F$  of a size  $|A| \times |A|$ , such that

$$F_{i,j} = \begin{cases} f_{i,j} & \text{if } i \neq j, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Similarity of feature sets that co-occur with  $a_i$  and  $a_j$  can be computed, e.g., as a cosine between the corresponding rows of matrix  $F$ , i.e.:

$$simil(F_{i,\cdot}, F_{j,\cdot}) = cosine(F_{i,\cdot}, F_{j,\cdot}) = \frac{\sum_{l=1}^{|A|} F_{i,l} \cdot F_{j,l}}{\|F_{i,\cdot}\| \cdot \|F_{j,\cdot}\|} \tag{3}$$

The final value of the proposed interchangeability measure is a difference between the above similarity and the frequency with which features co-occur:

$$I(a_i, a_j) = simil(F_{i,\cdot}, F_{j,\cdot}) - F_{i,j} \tag{4}$$

The proposed measure has similar properties as the measure proposed in [5] for quantifying interchangeability of cards in collectible card game’s decks. It can take values from  $[-1, 1]$  and the value 1 is taken only if the two compared features are perfectly interchangeable (in particular  $I(a_i, a_i) = 1$ ). Values close to  $-1$  are taken only in case when one feature is present only in a small number of reducts and is always accompanied by the second feature that also appears in many different reducts (the cosine between any two rows of  $F$  cannot be lower than 0). If  $a_i$  and  $a_j$  always appear in the same reducts, then  $I(a_i, a_j) = 0$ . Also, the measure is not symmetric. If two features never appear in the same reduct but the first one is much more relevant to the considered problem, it is much safer to exchange the second one for the first than the other way around.

One of possible applications of the proposed measure is to cluster statistics extracted from time windows into groups of pairwise interchangeable features. For this purpose,  $I$  can be transformed into a symmetric dissimilarity function by taking  $diss(a_i, a_j) = 1 - (I(a_i, a_j) + I(a_j, a_i)) / 2$ . A short comment is also needed to explain why it is our strong belief that approximate reducts computed with DAAR are a good choice for measuring such interchangeability. The main reason is that reducts obtained using DAAR are less likely to contain features that represent the aforementioned illusionary dependencies that could bias the frequencies of co-occurrences between pairs of features and introduce random noise to values of the matrix  $F$ . Moreover, DAAR adds diversity to a set of considered approximate reducts due to a fact that it allows to include feature sets that correspond to different approximation thresholds. As a result, it makes it easier to capture complex dependencies between features.



## 5 Evaluation of the Proposed Data Modeling Chain

We conducted a series of experiments on two data mining competition data sets mentioned in Sect. 2. In the first step, for each set we constructed features in accordance to the description provided in Sect. 3. This way, for the three considered methane-related tasks (corresponding to sensors *M256*, *M263*, *M264*) we obtained representation comprising of 2100 features, whereas for the seismic activity-related data we worked with 743 features. In order to compute rankings of constructed features using DAAR, we discretized the available training data sets using the local discernibility method implemented in *RoughSets* library. For each of the considered decision tasks, values of each feature were divided into four intervals. The number of intervals was chosen such that it reflected a typical ordinal scale used by experts during the assessment of mining hazards.

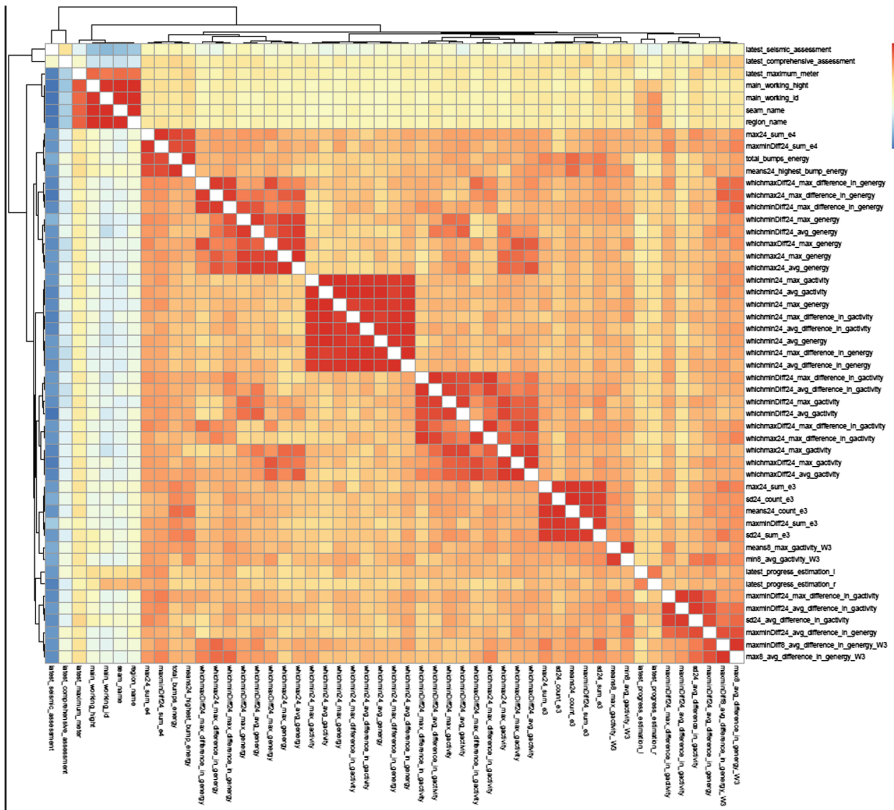
After discretization, for each task we used the training data to derive  $2^{13} = 8192$  approximate reducts. DAAR's version available in *RoughSets* was used for this purpose<sup>4</sup>. The resulting sets of reducts were quite diverse. For instance, for the seismic activity data, their length ranged from 4 to 13 with a median equal to 9. Interestingly, only 427 features appeared at least once in reducts.

The computed reducts were used to create feature rankings. As expected, the most important features related to predicting seismic activity were those corresponding to expert's assessments of safety conditions, e.g., the feature *latest\_seismic\_assessment* was present in more than 50% of all computed reducts. This finding confirms validity of assessment approaches routinely applied in Polish mines. The *seismic\_assessment* method is based on the analysis of a registered frequency and energy of rock-bursts. The *comprehensive\_assessment* that corresponds to the second feature in the ranking (*latest\_comprehensive\_assessment*), is a combination of several other methods used by experts (e.g.: seismic and seismoacoustic). Other variables that turned out to be important correspond to the activity of geophones and the speed of the mining process. Their presence among top-ranked features is fully justifiable, since they are based on the same sensor types as the expert methods. However, the ranking reveals which types of aggregations of sensor readings bring more useful information to predictive models and thus, can be used to improve over the expert methods.

For the methane-related data, in prediction of dangerous concentrations at methanometers *M256* and *M264* the key features were related to max values of the most recent readings of those sensors. Interestingly, among five top-ranked features for *M263* only two were directly related to readings of that sensor – the three other features referred to readings of an anemometer, a barometer (atmospheric pressure) and to recent changes in temperature readings. For this methanometer, a key factor related to the prediction of methane concentration levels was the ventilating intensity of the whole longwall area (sensor *AN311*). This finding was quite unexpected because readings of *AN311* and their simple aggregations do not correlate well with methane levels by themselves. However,

---

<sup>4</sup> We kept most of the settings as their defaults – the only two modified parameters were *allowedRandomness* fixed at 0.05 and *semigreedy* set to *TRUE*.

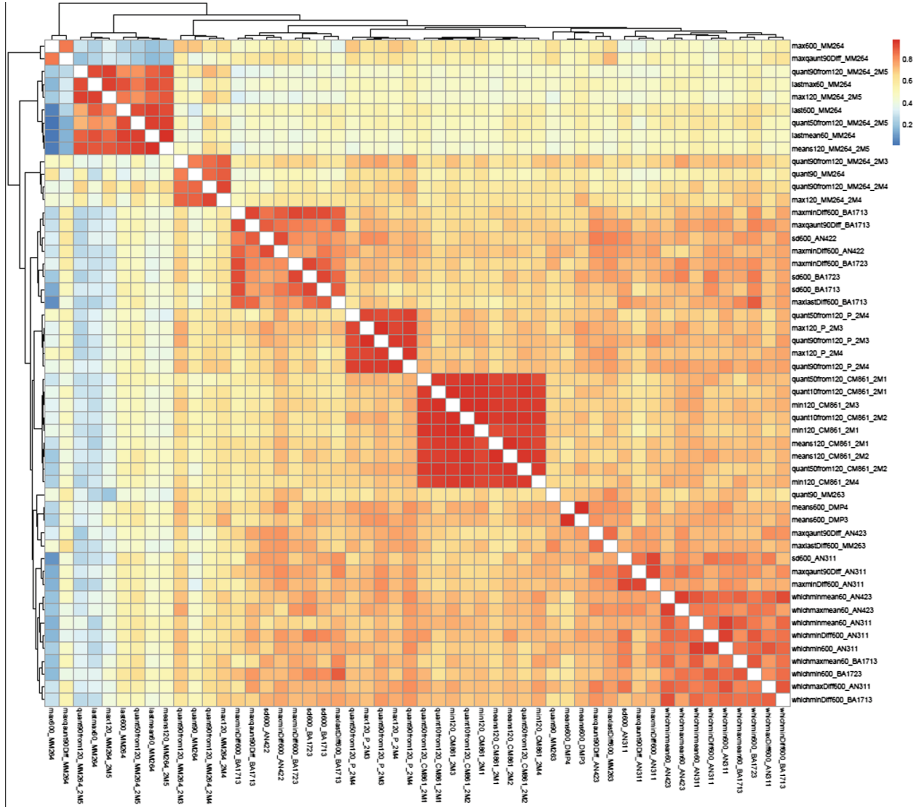


**Fig. 2.** Interchangeability of the most important features related to prediction of increased seismic activity periods. The warmer the color is, the more interchangeable are the features at the corresponding row and column.

added to a logistic regression model that operates only on readings from *M263*, they significantly improved its average AUC of predictions. Experts who were asked to explain this phenomenon stated that it is most likely justified by a specific location of the *M263* sensor at an edge of the longwall, in an area where a turbulence of air and methane may occur due to excessive ventilation.

In the next step, for each task we created the interchangeability matrices as it was described in Sect. 4. These matrices, after a transformation into feature dissimilarity values, served as input for a hierarchical clustering algorithm<sup>5</sup>. Figures 2 and 3 show interchangeability and clustering visualizations of the most important features for the seismic and *M264*-related case studies. These features correspond to top *k* features from the corresponding rankings, where *k* was selected for each task using a typical permutation test.

<sup>5</sup> An implementation of hierarchical clustering with complete linkage function from the *cluster* library of R system was used.



**Fig. 3.** Interchangeability of the most important features related to prediction of methane concentrations at methanometer *M264*. The warmer the color is, the more interchangeable are the features at the corresponding row and column.

In case of features related to seismic activity, noticeable is sharp separation of those related to expert assessments of safety conditions, those describing particular working sites and those corresponding to occurrences of latest min/max values of aggregated *genergy* and *gactivity* measurements. For the data related to *M264*, the most distinctive groups of features are those indicating recent readings of the target methanometer and those which show how high were these readings during a period between fourth and eighth minute of a time window. Apparent is also a group of features related to one of anemometers and a barometer, as well as a group expressing max values of pressure inside the methane drainage pipeline. Finally, there is also a visible group of features that show min values of methane readings captured in first two minutes of a window.

Apart from indicating similar features, the interchangeability analysis provides useful knowledge regarding attributes that should not be exchanged by others. These features could be seen as indispensable in constructing reliable prediction models. For example, the expert assessments of safety are difficult to

replace by a small set of different features. In a wider perspective, information about such features can be of high importance. Maintenance services should take adequate care of sensors that provide unique information. In case of failure of such a device, predictions might be imprecise or even impossible. This refers not only to natural hazards but also to other domains, such as machine monitoring (condition monitoring, preventive maintenance, etc.). Moreover, information about feature interchangeability can be directly used in a construction of prediction models. One example of such an application is demonstrated below. It is based on an idea of using a small and diverse set of features to build a single efficient model that can be easily interpreted by subject matter experts. In a different approach one may use many sets of non-interchangeable features to construct an ensemble of prediction models. However, since we focus on model's interpretability and informative representations of time series data, in further investigation we confine ourselves to the use of a single feature subset.

As already discussed, from a perspective of practical applications of predictive analytics in the mining industry, an important aspect of any model is its interpretability. Since all decision support and hazard monitoring systems are overseen by human experts, their results have to be comprehensible and easy to explain based on available data and domain knowledge. For this reason, relatively simple methods, such as rule-based classifiers or linear models, are very appealing in this application area. In this section we concentrate on the latter of those two approaches, namely on the logistic regression [4].

In order to construct interpretable predictive models for our case studies, we decided to incorporate the obtained results of feature clustering into a final feature selection process. Basically, we choose a single feature from each group on a basis of the previously prepared rankings – in each cluster we select a feature with the highest rank. However, for this procedure to work it was necessary to produce an appropriate number of feature clusters. In our experiment, we did it automatically using the wrapper approach. We checked divisions into consecutive numbers of features groups (between 2 and 15 groups), defined by the hierarchical clustering trees computed for each data set. For each division, we selected the best feature in each group and trained 1000 logistic regression models on different stratified samples of training data cases (the feature subset was fixed). Due to highly imbalanced occurrence of target labels, we sampled with repetitions from the ‘*warning*’ classes and added the same number of cases from the ‘*normal*’ classes, so that the total number of training cases for each logistic regression model was twice as large as the number of ‘*warning*’ cases in the training data. A standard implementation of logistic regression from *base* R was employed for model training, without any additional regularization.

The learnt models were assessed on out-of-bag sample of the training data (AUC measure was used on probabilities returned by models) and the average scores were recorded for divisions into different amounts of clusters. Additionally, models made their predictions on the test data and their average preliminary

evaluation scores were computed<sup>6</sup>. The final number of clusters, and thus the number of features used by logistic regression, was decided based on average of those results. When the number of features was fixed, an ensemble of the 10 best of the corresponding models was taken as the final classifier.

Tables 2 and 3 illustrate a comparison between performance of the resulting model and the top-ranked approaches from two data mining competitions. We performed 10 independent repetitions of our model building process and we report the mean and standard deviation of our outcomes. Since results of the ensemble were based on random data sampling, the final number of chosen features for each task was slightly different in each run of the experiment.

**Table 2.** Final and preliminary results of the top seven from 52 actively participating teams in challenge [15], in comparison to results of the described ensemble of logistic regression models combined with our feature selection method. In case of our model, the average result and its standard deviation is given from 10 independent runs.

Rank	Team name	Preliminary	Final score
1	zagorecki	0.9666	0.9593
–	our model	$0.9398 \pm 0.0078$	$0.9559 \pm 0.0035$
2	marchb	0.9461	0.9439
3	dymitrruta	0.9337	0.9437
4	moomean	0.9286	0.9428
5	trzewior	0.9469	0.9413
6	kkurach_kp7	0.9685	0.9400
7	tozczypaczek	0.9484	0.9398

**Table 3.** Final and preliminary results of the top seven from 106 actively participating teams in challenge [6], in comparison to results of the described ensemble of logistic regression models combined with our feature selection method. In case of our model, the average result and its standard deviation is given from 10 independent runs.

Rank	Team name	Preliminary	Final score
1	tadeusz	0.9200	0.9393
2	deepsense.io	0.9338	0.9384
3	yata	0.9161	0.9342
–	our model	$0.9469 \pm 0.0001$	$0.9337 \pm 0.0006$
4	podludek	0.9304	0.9336
5	jellyfish	0.9320	0.9336
6	millcheck	0.9151	0.9329
7	kkurach	0.9316	0.9312

<sup>6</sup> It is worth to realize that the preliminary scores were available to participants during the competitions and they were commonly used to tune parameters of solutions.

**Table 4.** A comparison of results obtained for different feature selection (FS) methods using the same prediction model. Average AUC and standard deviations are given.

FS method	Seismic activity		Methane ( <i>M263</i> )		Methane ( <i>M264</i> )		Methane ( <i>M256</i> )	
	Score	N.of. feat.	Score	N.of. feat.	Score	N.of. feat.	Score	N.of. feat.
Our model	0.9337 ± 0.0006	8	0.9621 ± 0.0033	7	0.9556 ± 0.0009	4	0.9499 ± 0.0004	9
AIC	0.9250 ± 0.0006	31	0.9044 ± 0.0050	26	0.9512 ± 0.0004	17	0.9465 ± 0.0008	14
Correlation filter	0.9210 ± 0.0010	70	0.8604 ± 0.0143	94	0.9555 ± 0.0007	64	0.9462 ± 0.0012	62
Greedy reduct	0.9147 ± 0.0007	8	0.9527 ± 0.0022	12	0.9524 ± 0.0008	8	0.9567 ± 0.0006	10
<i>mRMR</i>	0.9067 ± 0.0007	4	0.9445 ± 0.0010	6	0.9504 ± 0.0003	4	0.9503 ± 0.0006	9
Correlation clusters	0.9227 ± 0.0007	19	0.9457 ± 0.0032	15	0.9490 ± 0.0013	19	0.9456 ± 0.0020	7
No FS – baseline	0.8785 ± 0.0012	743	0.8129 ± 0.0021	2100	0.8129 ± 0.0016	2100	0.8129 ± 0.0015	2100

However, the differences were small and they did not have much impact on the final evaluation, as confirmed by low standard deviations. In the most number of runs, for seismic data only 8 features were used, and for the methane-related tasks only 9, 7 and 4 features were chosen for methanometers *M256*, *M263* and *M264*, respectively. Noticeable is the fact that even though our ensembles took into account only a very small number of different features, they were able to outperform the most of other, often significantly more complex approaches (e.g.: SVMs, random forests, gradient boosting trees, deep learning methods, etc.).

In order to additionally confirm soundness of our feature selection method, we decided to compare the results of our model to those obtained with the same prediction method but using different algorithms for choosing the final feature sets. We computed feature subsets using the aforementioned *mRMR*, a correlation-based filter with a permutation test for determining the number of selected features, as well as a greedy search based on AIC criterion that is commonly applied to select a logistic regression model [2]. For each data set, we also computed a single approximate reduct using DAAR. Finally, we tried a different approach to feature clustering, in which the similarity between features corresponds to absolute value of their linear correlation. Table 4 contains the outcomes of all the compared feature selection algorithms.

A significance of differences between results of different feature selection methods was verified using Wilcoxon rank sum test. For seismic hazard predictions, our method significantly outperformed all other algorithms at 0.99 confidence level. Moreover, based on the performed analysis of feature interchangeability our model selected only 8 different features. Among other tested approaches *mRMR* yielded a smaller set (only 4 features) but in that case the obtained performance was considerably weaker. In fact, AUC for *mRMR* method was lower than for predictions made by experts. Another method that yielded the same number of features was DAAR, though the corresponding results were comparable to those achieved by the expert method.

For predicting methane concentration, our feature selection method denoted the highest average of AUC scores from the three methanometers (0.9559) as well. For this task, however, the results obtained by other algorithms were much closer. In fact, for predictions of warnings at methanometer *M256* the best set of features was selected by computing the DAAR-based approximate reduct (the difference in results with our method was statistically significant). For this sensor, slightly better evaluation was also obtained using *mRMR* but in that case the difference with our approach was not determined as significant ( $p\text{-value} > 0.05$ ). For the remaining two methanometers, average results obtained on features indicated using our method were always better than other approaches.

## 6 Conclusions

We described a methodology for constructing concise, yet informative representations of time series of sensor readings. Such representations can be utilized for constructing interpretable prediction models in a context of monitoring mining hazards in underground coal mines. In our case study, we demonstrated that the proposed heuristic for measuring interchangeability of features describing sensory data streams can facilitate selecting small feature subsets that are sufficient to robustly detect periods of increased seismic activity and foresee warning levels of methane concentrations. In conducted experiments, our simple models were able to outperform many complex prediction algorithms and could successfully compete with top-ranked approaches from the related data mining competitions that we organized on the Knowledge Pit platform.

The analytic process demonstrated in this paper can be applied to a wide array of problems, not necessarily related to underground coal mining. In fact, the approach is independent of particular sensor types, which was confirmed in our experiments – the two considered case studies significantly differed in types and characteristics of sensors used for gathering the data. Analogous solutions can be considered for other problems specific to mining and excavatory operations, with little impact of their nature (deep or open-pit mines, type of material, background seismic activity, etc.). In particular, our approach could be used to investigate and control the environmental impact of open-pit mines, mitigate risks and improve hazard prevention procedures, and to facilitate predictive maintenance and cost efficiency optimization of mining operations.

In the future, we plan to extend the described framework by an aspect of interactions with subject matter experts in the process of detecting interchangeable features and selecting an optimal feature subset. Incorporation of additional domain knowledge could greatly improve the feature clustering results. One way of achieving this is to combine semi-supervised modeling with the active learning approach. In this way, we would be able to smartly select pairs of features and obtain for them information regarding their similarity, e.g., in a form of must-links and cannot-links. Not only such links could be directly employed for the feature selection but could also provide means for designing fail-safe systems for monitoring and prevention of mining hazards.

## References

1. Ellenberger, J.L., Heasley, K.A.: Coal mine seismicity and bumps: historical case studies and current field activity. In: Proceedings of ICGCM 2000, pp. 112–120 (2000)
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)
3. He, M., Sousa, L.R., Miranda, T., Zhu, G.: Rockburst laboratory tests database: application of data mining techniques. *Eng. Geol.* **185**, 116–130 (2015)
4. Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, 3rd edn. Wiley, Hoboken (2013)
5. Janusz, A., Grad, L., Ślęzak, D.: Utilizing hybrid information sources to learn representations of cards in collectible card video games. In: Proceedings of ICDM 2018 Workshops, pp. 422–429 (2018)
6. Janusz, A., Grzegorowski, M., Michalak, M., Wróbel, L., Sikora, M., Ślęzak, D.: Predicting seismic events in coal mines based on underground sensor measurements. *Eng. Appl. Artif. Intell.* **64**, 83–94 (2017)
7. Janusz, A., Ślęzak, D.: Computation of approximate reducts with dynamically adjusted approximation threshold. In: Proceedings of ISMIS 2015, pp. 19–28 (2015)
8. Janusz, A., Ślęzak, D.: Rough set methods for attribute clustering and selection. *Appl. Artif. Intell.* **28**(3), 220–242 (2014)
9. Kabiesz, J., Sikora, B., Sikora, M., Wróbel, L.: Application of rule-based models for seismic hazard prediction in coal mines. *Acta Mont. Slovaca* **18**(4), 262–277 (2013)
10. Kruczyk, M., Baltzer, N., Mieczkowski, J., Damiński, M., Koronacki, J., Komorowski, J.: Random reducts: a Monte Carlo rough set-based method for feature selection in large datasets. *Fundam. Inform.* **127**(1–4), 273–288 (2013)
11. Moczulski, W., Przystałka, P., Sikora, M., Zimroz, R.: Modern ICT and mechatronic systems in contemporary mining industry. In: Proceedings of IJCRS 2016, pp. 33–42 (2016)
12. Pawlak, Z., Skowron, A.: Rough sets and Boolean reasoning. *Inf. Sci.* **177**(1), 41–73 (2007)
13. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1226–1238 (2005)
14. Riza, L.S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Ślęzak, D., Benítez, J.M.: Implementing algorithms of rough set theory and fuzzy rough set theory in the R package ‘roughsets’. *Inf. Sci.* **287**, 68–89 (2014)
15. Ślęzak, D., Grzegorowski, M., Janusz, A., Kozielski, M., Nguyen, S.H., Sikora, M., Stawicki, S., Wróbel, L.: A framework for learning and embedding multi-sensor forecasting models into a decision support system: a case study of methane concentration in coal mines. *Inf. Sci.* **451–452**, 112–133 (2018)
16. Wróblewski, J.: Ensembles of classifiers based on approximate reducts. *Fundam. Inform.* **47**(3–4), 351–360 (2001)