INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 3(33) • 2014
ISSN 1507-3858

Bartłomiej Karaban, Jerzy Korczak
Wrocław University of Economics
e-mails: {bartlomiej.karaban; jerzy.korczak}@ue.wroc.pl

SELECTION OF ATTRIBUTES FOR A CLASSIFIER OF TELECOMMUNICATION FAILURES IN THE COPPER MINE[1]

Abstract: Ensuring safety and continuity of production is the major task of telecommunication systems in deep mining. These systems, despite their use of modern and innovative infrastructure of monitoring solutions, are not free from imperfections. One of the practical problems are false alarms signaling the occurrences of damaged infrastructures. In the paper, the data sources of the telecommunication system are identified and described, as well as the methods of their preprocessing. To build a classifier, a method of attribute selection is proposed to detect false alarms generated by the telecommunication system of the mine. Experiments were carried out on real data extracted from the telecommunication system operating in the copper mine of the KGHM Polska Miedź SA.

Keywords: Data mining, telecommunication system, data preprocessing, classification methods, inductive decision trees, detection of false alarms, decision rules.

DOI: 10.15611/ie.2014.3.06

1. Introduction

Work safety in the mine, good organization, and continuity of production require an efficient and effective monitoring system of the state of the telecommunications installation, machinery, equipment, and employees. One of the important functions of a monitoring system is to collect and provide operators with information about the status of communication with network devices, the presence of voltage current, time, location of failures, and values of parameters of the equipment. A failure is understood by the operators as a loss of connectivity with a particular device [Karaban 2013].

This article presents a study on a well-known and difficult problem occurring in monitoring systems - the problem of selecting attributes to identify false alarms. The study has been conducted in the mining companies belonging to KGHM Polska

[1] Selected parts of this article were published under nonexclusive copyright in Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS 2014 (see [Karaban, Korczak 2014]).

Miedź SA. It should be noted that the monitoring systems in the mine always generate an alarm about the lack of communication in cases of actual loss.

The purpose of this publication is to present an approach to data preprocessing and attribute selection to correctly recognize false alarms generated by the monitoring system, and finally to extract new and useful rules of identification, in particular the false alarms.

In the following section the data will be detailed. Section 3 presents the method of data exploration. The results of experiments are discussed in section 4. The experiments carried out on real data extracted from the system database are also presented.

2. Description of the data

The main source of data was the database of the system that controls the telecommunication infrastructure in the mine. The system consists of the physical components and software packages that within a data communication system allow one to carry out the telecommunication functions, administrative services and diagnostics. The architecture and functionalities of the system have been described in [Korczak, Karaban 2014].

The analyzed data were real data extracted from: system monitoring, technical documentation, and system entries in the service log and saved in the Monitoring Database. The data set was randomly selected from 01.05.2012 to 18.05.2012. In a first stage, necessary transformations of attribute values were carried out in order to use the data in the selected classification algorithms. The length of a 2.5-week time series is justified in terms of representativeness of the sample: it has every work shift, every day of the week and days when the mine does not conduct mining. The drawn data set contained 1481 *n*-tuples, of which 89% (1316) were alarms about the lack of communication. Each alarm was described with the following attributes:

- *device* – the identifier of the device, which is in the state of alarm; it can be used as an identifier of the observation,
- *type* – the variable describing the device by the performed function,
- *description* – the information about the localization of the device and its physical address,
- *status* – the state of the parameter of the variable to which the alarm applies (for example, “absence” for the communication or the supply voltage, “very high” or “high” for the incoming current),
- *confirmed* – time when the occurring alarm was generated (yyyy-mm-dd hh:mm:ss),
- *date of retreat* – time when the alarm was stopped,
- *alarm duration* – the duration of the alarm (hh:mm).
- *value* – the numerical value assigned to the variable such as, for example: incoming current, outgoing current, supply voltage.

Attributes: *device*, *description*, *status*, *value*, *confirmed* are not included, from the operational viewpoint of information that could be useful to extract a new knowledge, therefore they were considered in modeling, however they can serve as descriptive attributes that might be helpful in interpretation of results. It should be pointed out that the reduction of the data space decreases the risk of excessive fit of the model, and decreases the rule complexity, and ipso facto increases the quality of rule generalization [Zhang et al. 2011].

In this step, the raw set of data was also filtered by the criterion of practical significance. However the preselected list of attributes was insufficient to extract new operational knowledge. For instance, the only relevant attribute chosen to recognize a false alarm was short time of duration, wherein were unknown the lower and upper values when the alarm would be considered false. Similarly, some devices generated more alarms, but there was no detailed analysis available to insert this information into the model.

In order to check if the false alarms can be recognized among the others, the new attributes have been added, such as a type or subtype of device, district station, a district of mining, a branch or branch type. The values of these attributes were manually entered into a new table based on the information extracted from the telecommunication system and the maps of the telecommunications infrastructure.

Due to the use of methods of supervised classification, it was necessary to determine the value of the target class *alarm*. The class takes two values: *true alarm* or *false alarm*, given according to the historical alarms and the service book.

To create an input file for the preliminary analysis, and then for alarm recognition and classification, all data extracted from the technical documentation, maps of the infrastructure and service book and the database diagnostic system had to be integrated.

After this short introduction and the presentation of data description, the next section presents the most important transformation of data and the criteria used to select attributes for modeling.

3. Data preprocessing and attribute selection

The heterogeneity of the data and the specificity of classification algorithms require a number of preprocessing operations before building of the classifier. For example, the CN2 algorithm requires discrete data (qualitative) as an input. Decision tree algorithms can use the continuous attributes, but in general, the use of such attributes results in an increase in the cost of computation and the complexity of the generated rules. Discretization, however, increases the interpretability of generated hypotheses, especially in the case of not completely correct training data [Sang et al. 2014; Madolando et al. 2014].

In the analyzed case, three attributes: *alarm duration*, *date of occurrence*, and *date of retreat* were transformed. The format of the attributes was transformed into the format (hh:ss), then recorded as numerical values in an Excel spreadsheet.

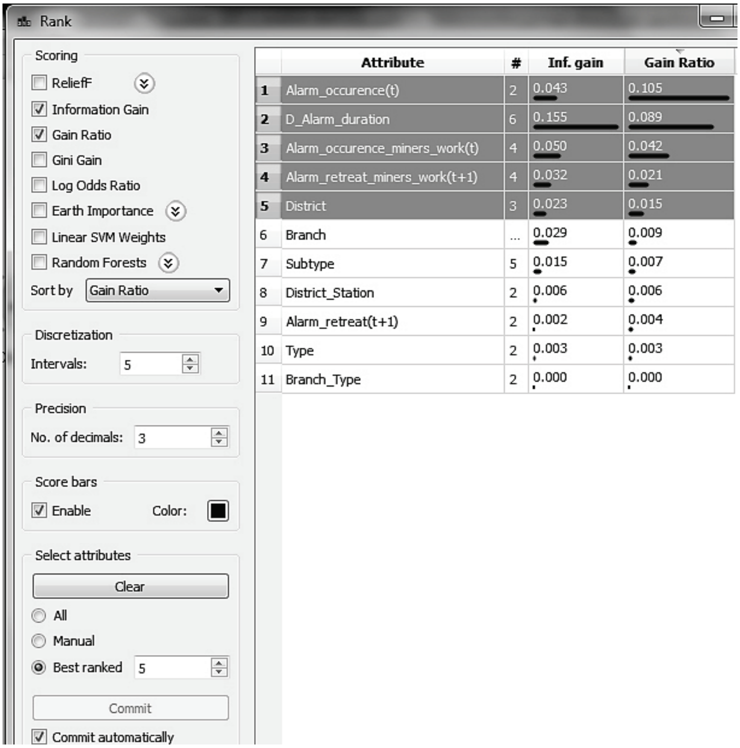
In the literature, many discretization methods are described. The simplest (sometimes called primitive) is a method of constant width classes and equal density (frequency) classes. A slightly more advanced one is the method MDL (Minimal Description Length) based on the theory of information [Korbicz et al. 2002].

These methods are not free from drawbacks arising from the nature of the problem analyzed and the distribution of a quantitative variable. Based on the statistics of the distribution of the variable *D alarm duration*, the choice of the method of equal width classes was not a good option (due to the presence of outliers and strongly asymmetric distribution of the target class). This might lead to a situation in which some classes would include a small number of alarms. In this case, a more appropriate method was the use of equal density classes, but this was associated with the risk of identifying the limits of classes of values, which might negatively influence the results of discrimination alarms. In view of the disadvantages, the most attractive alternative was the MDL discretization. Furthermore, the rules having in the condition the duration indicate that it may be considered as a good attribute to identify a false alarm, but not in the attempt to discover new, useful knowledge. Taking into consideration the above arguments, in the project, equal density classes, with manual tuning "pure" classes, have been used in the discretization method.

Another transformation was made to assign the values of variables *date of occurrence* and *date of retreat*. In order to obtain the most general rules, each of these attributes is represented by two other attributes of the different levels of *granulation* of information. The newly created attributes form the following attributes: *date of occurrence* is *alarm_occurrence* (*t*) and *alarm_occurrence_miners_work* (*t*), by analogy with the attribute *date of retreat*, *alarm_retreat* (*t* + 1) and *alarm_retreat_miners_work* (*t* + 1). In the cases of attributes *alarm_occurrence* (*t*) and *alarm_retreat* (*t* + 1) the time value was assigned to the *mining* or *stoppage* that carry out information about the occurrence of the alarm. As a result of the second transformation, the following attributes were obtained: *alarm_occurrence_miners_work* (*t*) and *alarm_retreat_miners_work* (*t* + 1) that can take values of *miners_relay*, *miners_overlay*, *blasting_works* or *stoppage*. These transformations were verified whether, among the generated rules, there exist conditions relating to the work intensity of people which might cause the traffic growth in the telecommunication network.

In the literature, many algorithms have been described to attribute selection. R. Kohawi and G.H. John propose the filtering algorithms independent of a used classifier, based on variance, the Pearson correlation, Fisher statistics, Chi-square statistics, Information Gain, and Gain Ratio [Kohavi, John 1997].

Most of these algorithms (e.g. based on the variance, correlation, or Fisher statistics) have not been used in the experiments due to the fact that we use the classification method for which the input data must be symbolic. Taking this into account, the data were transformed using the ranking criterion Gain Ratio. An additional advantage is that this criterion does not favor the attributes with a large number of values, as happens with the Information Gain [Zhang et al. 2011]. In the selection of attributes, the highly-ranked attributes were those with high descriptive values. Furthermore, for clarity of the rules, the number of selected attributes was arbitrarily limited to 6. The ranking result is shown in Figure 1.



Scoring
☐ ReliefF
☑ Information Gain
☑ Gain Ratio
☐ Gini Gain
☐ Log Odds Ratio
☐ Earth Importance
☐ Linear SVM Weights
☐ Random Forests
Sort by: Gain Ratio

Discretization
Intervals: 5

Precision
No. of decimals: 3

Score bars
☑ Enable Color: ■

Select attributes
Clear
○ All
○ Manual
● Best ranked 5
Commit
☑ Commit automatically

	Attribute	#	Inf. gain	Gain Ratio
1	Alarm_occurence(t)	2	0.043	0.105
2	D_Alarm_duration	6	0.155	0.089
3	Alarm_occurence_miners_work(t)	4	0.050	0.042
4	Alarm_retreat_miners_work(t+1)	4	0.032	0.021
5	District	3	0.023	0.015
6	Branch	...	0.029	0.009
7	Subtype	5	0.015	0.007
8	District_Station	2	0.006	0.006
9	Alarm_retreat(t+1)	2	0.002	0.004
10	Type	2	0.003	0.003
11	Branch_Type	2	0.000	0.000

Figure 1. The result of the ranking of attributes

After a series of experiments, the five attributes have been selected. Their choice determined not only the result of the ranking and the experiments carried out, but also the experience and expertise gained. Finally, the following attributes were selected:

- *Alarm_occurrence* (*t*), *Alarm_occurrence_miners_work* (*t*), *Alarm_retreat* (*t* + 1), because of the conditions identified during the preliminary analysis of data concerning false alarms;
- *Alarm_duration*, due to the possibility of establishing an objective, approximate threshold for the duration of the false alarm and the high position on the ranking results;
- *District*, due to the specificity of the districts and volume of traffic.

The next section describes a model, justifies the choice of classifier, and describes the results of experimentation.

4. Modeling and experimental results

G. Piatetsky-Shapiro defines the process of data mining and knowledge discovery as "the process of nontrivial extraction of potentially useful and previously unknown information, or general patterns existing in databases" [Piatetsky-Shapiro, Frawley 1991].

Referring to the specific problem of recognition and classification of alarms, it is possible to use two approaches well described in the literature: supervised or unsupervised classification. The main difference between them is *a priori* knowledge about the target class. Due to the possession of information whether a given historical alarm was true or false, it was desirable to apply a supervised classification.

There are many supervised classification methods which include: logistic regression, discriminant analysis, neural networks, genetic algorithms, SVM, Naive Bayes, CN2, and inductive decision trees. It is not possible to use many of these methods, for they are dedicated for the quantitative variables, while the alarm is mostly described using symbolic attributes. In contrast, methods such as SVM, Naive Bayes, and inductive decision trees do not have this restriction, therefore these methods were considered to be applied [Rivas et al. 2011].

Many experiments and tests of several methods for classification were carried out in the project, amongst which we identified inductive decision trees. For choosing this method, several conditions appealed. The first condition resulted from a fundamental principle of inductive inference leading to the generalization of observations and facts in the form of rules and statements. An analyst who has domain knowledge should verify the authenticity of the generated rules and model the tree, as long as these rules are pragmatic enough to apply them to solve the problem. Another important advantage was the simplicity of interpreting the derived rules in both graphical and decision rules. The last advantage, which prompted us to use tree induction, is the ability to control the complexity and generality of generated rules. The weakness of the IDT is the possibility of generating too large and too deep a tree which might overfit and generate erroneous classifications [Bramer 2013].

There are many measures for evaluating classifiers such as sensitivity, precision, specificity or accuracy [Morzy 2013]. The studies demonstrate that the most important objective is to assure that all true alarms will be identified, then the number of false alarms would be minimized. In our approach, the discovery will be focused on the minimization of the error of the first kind (False Positive rate), specifying the number of false alarms which were classified as true. The second important quality measure in this project was the precision, which takes into account the number of real alarms misclassified as false. In general, when choosing a classifier, we strive to achieve a compromise between readability and usability rules and maximizing the value of these measures. The first criterion is considered more important in the case of a small difference in the assessment of classification.

The process of data mining has been carried out according to the methodology CRISP-DM [Ding 2013], using the data mining platform Orange. The schema of the data mining process is shown in Figure 2. The first step, the data preprocessing, was performed (highlighted by orange color), part of which was done using MS EXCEL and MS ACCESS. The next step concerned the data analysis (box of brown contour), after which the model was built and applied to explore the data using the chosen methods of classification (box with a green outline). The last step was to evaluate the selected models (box with blue outline).

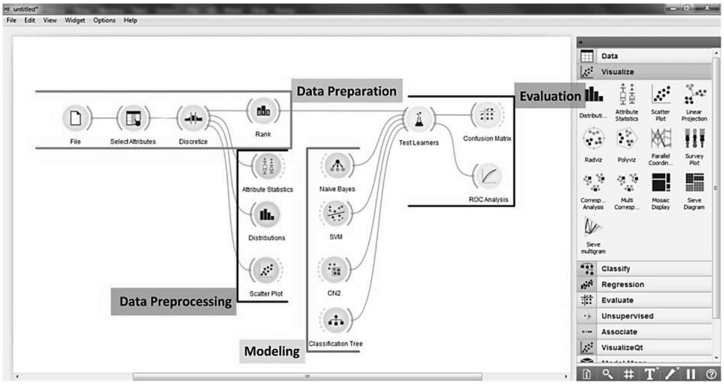


Figure 2. Schema of data mining process using Orange platform

In order to recognize and classify false alarms, a two-step approach was carried out. In the first step, the rules that classify the true alarms with the greatest possible purity were discovered. In Figure 3 this set of rules is named RulesTP. The algorithm is the following:

```
let the RulesTPSet be empty
  repeat
    Find_BestRule(TrueAlarm)
    if the BestRule is not nil
      then
        let the TrueAlarmsDiscovered be the observations
          covered by the BestRule
        remove from the TrueAlarms the observations in the
          TrueAlarmsDiscovered
        append the rule to the RulesTPSet
  until the TrueAlarms is empty or the BestRule is nil
return the RulesTPSet
```

Figure 3. Text of rules named RulesTP

After inducing rules that cover all true alarms, in the second step the rules to discover false alarms were generated.

In Figure 5, two sets of false alarms are shown; the blue one illustrates the false alarms indicated by the current system, but the rose one shows the false alarms generated by the RulesTP.

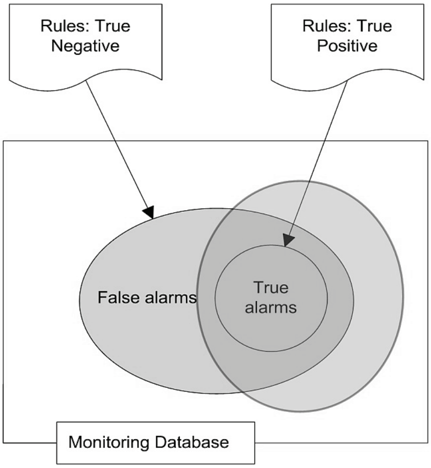


Figure 4. Schema of alarms identification by the rules

As indicated and justified at the beginning of the section, the decision tree induction algorithm based on entropy was chosen [Gorunescu 2011]. During the experiments, about 100 variants of decision trees were generated by changing the value of various parameters, such as the tree pruning ratio and the number of class attributes describing the duration of the alarm.

In the first series of experiments, the number of class attributes was specified: Alarm_duration to 10 (the maximum for the reduction in the package Orange). The controlled parameter was the minimum number of observations in the leaf, which value after 10 experiments was set to 5.

A tree was generated in which the 6 leaf nodes of the attribute Alarm_duration contained only observations of false alarms that were aggregated, which resulted in the reduction in their number from 10 to 5. The indicated change negatively affected the quality of all four classifiers. The obtained specificity ratio (TN) made the value worse than in previous experiments.

In the last series of experiments, the parameter to halt the tree construction was changed. Just as in the first series of experiments, during the test the number of observations in the leaf was monitored, at which point the construction of the tree was stopped. The values of the parameter ranged from 1 to 10. Satisfactory results in terms of measures of quality and usability of the generated rules were obtained for the trees with a minimum of two observations per leaf.

In about 100 experiments, hundreds of decision rules were generated. The article presents only four of the ten rules that provide useful, previously unknown knowledge about the false and true alarms generated by the monitoring system. One of these rules was the following:

```
IF
alarm_ occurrence(t) = ‘stoppage’
∧ alarm_retreat_miners_work(t+1)= ‘miners_overlay’ v ‘miners_relay’
THEN true_alarm
```

The indicated rule covers 39 cases of purity of 100%. After analyzing the cases, it turned out that the rule identifies alarms which arose as a result of switching off the electrical switchboard during the weekend. Application of this rule increased besides the efficiency of alarm recognition also indications of the switchgears which due to the communication requirement should be maintained in a continuous operation. This information was forwarded to the electrical services at the mine.

The next rule that identified the true alarm was more complex, namely:

```
IF
alarm_occurrence(t) = ‘mining’
∧ alarm_duration> 5 hours and 30 minutes
∧ alarm_occurrence_miners_work(t) = ‘blasting_work’
THEN true_alarm
```

The rule allows us to identify the cases where the loss of communication occurs most likely as a result of conducting blasting work. This rule covers only three cases (the purity of leaf 100%); however, taking into account the information it provides, we can identify the situation in which there is damage to the telecommunication line because of blasting work. Providing such information to the service may reduce the incidence of such situations; it helps to reduce system failures and the costs associated with damage to the network. The duration of alarm in the last rule may seem too long. Such situations are due to the specific nature of the work environment in the mine (climatic and geological conditions), work organization (logistics, time to arrive at the place of an accident), and the difficulties encountered in removing failure (signs of non-access to dangerous places).

Having analyzed the generated rules, it was noted that in each of the created trees after the aggregation intervals of Alarm Duration (the third series of experiments) shows the strong rule purity (96.5%), covering ca. 800 alarms:

```
IF
alarm_occurrence(t) = ‘mining’
∧ alarm_duration ≤ 45 minutes
THEN false_alarm
```

This rule is a useful new piece of knowledge about the alarm time threshold below which an alarm can be considered with high probability to be false. The value of this threshold is approximately 45 minutes.

The next rule classifies the alarms of long duration:

```
IF
alarm_occurrence(t) =’mining’
∧ alarm_duration > 5 hours and 30 minutes
∧ alarm_occurrence_miners_work(t) = ‘miners_relay’
∧ alarm_retreat(t+1) = ‘stoppage’
THEN false_alarm
```

It should be noted that in practice it is particularly difficult to determine the authenticity of such alarms.

The rules given provide useful knowledge about a false alarm of long duration. As indicated in the introduction, operators often use heuristics when trying to determine the veracity of a given alarm. So far it has been assumed that the false alarm takes no more than two hours. Based on the first rule, it can be induced that in 45 cases the alarms of about 5.5 hours duration were false alarms (purity of the leaf = 100%), which exemplifies the existing heuristic assumptions.

5. Conclusions

Monitoring of the telecommunication infrastructure in the mining industry plays a key role in terms of safety, good organization, and continuity of production. Despite modern and innovative solutions, these systems are not free from drawbacks. The paper proposes an approach to detect false alarms about the lack of communication in the mine. The data sources and the methods of transformation and selection of attributes to build the classifier were detailed. A new operational knowledge has been acquired, e.g. to identify individual devices which were deprived of the power supply in the days that mining is not carried out, recognition of devices that can be damaged by blasting work, or about the alarm duration threshold below which the alarm can be considered objectively false.

The preliminary results are sufficiently promising to merit undertaking further research on alarms recognition, generated by the monitoring system in the mine. To follow up, the studies should be conducted on a much larger number of observations. To improve the quality of classification, additional attributes should be considered. Nevertheless, the obtained results can be used to ameliorate the efficiency of the currently exploited monitoring system.

References

Bramer M., 2013, *Undergraduate Topics in Computer Science. Principles of Data Mining*, Springer, London.

Ding S., 2013, *Model-Based Fault Diagnosis Techniques. Design Schemes, Algorithms and Tools*, Springer, London.

Gorunescu F., 2011, *Data Mining. Concepts, Models and Techniques*, Springer, Berlin.

Karaban B., 2013, *Indukcyjne drzewa decyzyjne w analizie alarmów systemu telekomunikacyjnego*, Uniwersytet Ekonomiczny we Wrocławiu, Wrocław [master thesis].

Karaban B., Korczak J., 2014, *An approach to discover false alarms in monitoring system in the copper mine*, [in:] Ganzha M., Maciaszek L., Paprzycki M. (eds), *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*. *Annals of Computer Science and Information Systems*, vol. 2, Polskie Towarzystwo Informatyczne, Warsaw, Institute of Electrical and Electronics Engineers, New York City, pp. 307–312.

Kohavi R., John G.H., 1997, *Wrappers for feature subset selection*, Artificial Intelligence, vol. 97, pp. 273–324.

Korbicz J, Kościelny J., Kowalczuk Z., Cholewa W., 2002, *Diagnostyka procesów. Modele. Metody sztucznej inteligencji. Zastosowania*, WNT, Warszawa.

Korczak J., Karaban B., 2014, *Metoda wykrywania fałszywych alarmów w systemie monitorującym sieć telekomunikacyjną kopalni*, Przegląd Górniczy, nr 70, pp. 108–112.

Madolando S., Weber R., Famili F., 2014, *Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines*, Information Sciences, vol. 286, pp. 228–246.

Morzy T., 2013, *Eksploracja danych. Metody i algorytmy*, Wydawnictwo Naukowe PWN, Warszawa, pp. 326–327.

Piatetsky-Shapiro G., Frawley W., 1991, *Knowledge Discovery in Databases*, The AAAI Press, Menlo Park.

Rivas T., Paz M., Martín J.E., Matías J.M., García, J.F., Taboada J., 2011, *Explaining and predicting workplace accidents using data-mining techniques*, Reliability Engineering & System Safety, vol. 96, pp. 739–747.

Sang Y., Qi H., Li K., Jin Y., Yan D., Gao S., 2014, *An effective discretization method for disposing high-dimensional data*, Information Sciences, vol. 270, pp. 73–91.

Zhang K., Li Y., Scarf P., Ball A., 2011, *Feature selection for high-dimensional machinery fault diagnosis data using multiple models and Radial Basis Function networks*, Neurocomputing, vol. 74, pp. 2941–2952.

DOBÓR ATRYBUTÓW W BUDOWIE KLASYFIKATORA ALARMÓW SYSTEMU TELEKOMUNIKACYJNEGO KOPALNI

Streszczenie: Zapewnienie bezpieczeństwa pracy i utrzymanie ciągłości wydobycia to kluczowe zadania systemów telekomunikacyjnych w górnictwie głębinowym. Systemy te, pomimo nowoczesnych i innowacyjnych rozwiązań monitorowania infrastruktury, nie są wolne od wad. Praktycznym problemem jest występowanie fałszywych alarmów o uszkodzeniu infrastruktury. W publikacji wskazano źródła danych, opisano dane, metody ich przekształcenia oraz doboru zmiennych do budowy klasyfikatora. Następnie zaproponowano metodę wykrywania fałszywych alarmów w systemie telekomunikacyjnym kopalni oraz zaprezentowano niektóre reguły dostarczające użytecznej wiedzy z danych. Eksperymenty zostały przeprowadzone na rzeczywistych danych pochodzących z systemu telekomunikacyjnego funkcjonującego w kopalni KGHM Polska Miedź SA.

Słowa kluczowe: *data mining*, system telekomunikacyjny, drzewa indukcyjne, fałszywe alarmy, przygotowanie danych, reguły klasyfikacyjne.