Mining Data from Coal Mines: IJCRS'15 Data Challenge

Andrzej Janusz[1](✉), Marek Sikora[2], Łukasz Wróbel[2,3], Sebastian Stawicki[1,4],
Marek Grzegorowski[1], Piotr Wojtas[3], and Dominik Ślęzak[1,4]

[1] Institute of Mathematics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
janusza@mimuw.edu.pl, m.grzegorowski@mimuw.edu.pl

[2] Institute of Computer Science, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
marek.sikora@polsl.pl

[3] Institute of Innovative Technologies EMAG, Leopolda 31, 40-189 Katowice, Poland
lukasz.wrobel@ibemag.pl

[4] Infobright Inc., Krzywickiego 34, lok. 219, 02-078 Warsaw, Poland
{stawicki,slezak}@mimuw.edu.pl

Abstract. We summarize the data mining competition associated with IJCRS'15 conference – IJCRS'15 Data Challenge: Mining Data from Coal Mines, organized at Knowledge Pit web platform. The topic of this competition was related to the problem of active safety monitoring in underground corridors. In particular, the task was to design an efficient method of predicting dangerous concentrations of methane in longwalls of a Polish coal mine. We describe the scope and motivation for the competition. We also report the course of the contest and briefly discuss a few of the most interesting solutions submitted by participants. Finally, we reveal our plans for the future research within this important subject.

Keywords: Data mining competitions · Time series data · Attribute engineering · Feature extraction · High dimensional data.

1 Introduction

Coal mining requires working in hazardous conditions. Miners in an underground coal mine can face several threats, such as, e.g. methane explosions or rock-burst. To provide protection for people working underground, systems for active monitoring of a production processes are typically used. One of their fundamental applications is screening dangerous gas concentrations (methane in particular) in order to prevent spontaneous explosions [1]. Therefore, for that purpose the ability to predict dangerous concentrations of gases in the nearest future can be even more important than monitoring the current sensor readings [2].

Partially supported by the Polish National Science Centre – grant 2012/05/B/ST6/-03215 and by the Polish National Centre for Research and Development (NCBiR) – grant PBS2/B9/20/2013 in frame of the Applied Research Programs.

Y. Yao et al. (Eds.): RSFDGrC 2015, LNAI 9437, pp. 429–438, 2015.
DOI: 10.1007/978-3-319-25783-9_38

Typically, a monitoring system in a coal mine is conjugated with a power supply system and may automatically cut the power off if it detects any viable threat. Such action is necessary to prevent accidents. However, every break in the mining process results in costly losses for the mine. If the monitoring system could foresee dangerous concentrations of methane in the nearest future, dispatchers that control the mining process would be able to make some adjustments in order to decrease the transmission of the gas (e.g. by reducing the currents consumed by the cutter loader and thus lowering the intensity of the coal drilling which causes the release of methane from the soil). By doing so, it would be possible to avoid the necessity of shutting down the whole heavy machinery working in the mine. In this way, the monitoring system coupled with a decision support system could not only increase the safety of miners working underground, but also significantly reduce the overhead of the mining process.

From a data processing point of view, a decision support system which could aid in controlling the coal mining process requires efficient methods for handling continuous streams of data [3]. Such methods have to be able to handle large volumes of data from multiple sensors. They also need to be robust with regard to missing or corrupted data. Moreover, a good decision support system should be easy to comprehend by the experts and end-users who need to have access not only to its outcomes, but also to arguments or causes that were taken into account. A few practical studies have been already conducted with this respect, relying on rule-based models for predicting the methane level [4]. However, the literature on this important subject is still very scarce.

One of very few research initiatives in that field is DISESOR - a Polish national R&D project aimed at creation of an integrated decision support system for monitoring of the mining process and early detection of viable threats to people and equipment working underground. The system developed in the frame of DISESOR integrates data from different monitoring tools. It contains an expert system module that can utilize specialized domain knowledge and an analytical module which can be applied to make a diagnosis of the mining processes. When combined, these modules are capable of reliable prediction of natural hazards. The idea to popularize this topic among the data mining community by organizing an open data challenge originated within this project.

In IJCRS'15 Data Challenge we aimed to address the problem of active monitoring and prevention of methane outbreaks. We decided on the formula of an on-line and open to a whole data mining community competition due to the fact that it allows to conveniently review and test performance of the available state-of-the-art approaches. It is also an objective way of verifying the viability of not only the predictive models but also the whole analytic processes which include preprocessing methods, feature extraction, model construction and post processing of predictions (ensemble approaches). As the host we used the Knowledge Pit platform [5] which we designed to support the organization of data mining competitions associated with data science-related conferences.

The following sections give more details about the competition and its results. Section 2 briefly describes the above-mentioned Knowledge Pit platform and highlights its main functionalities. Section 3 explains our motivation to organize the competition and shows details of the utilized data set and the chosen evaluation method. Section 4 summarizes the competition results and gives some insights regarding the most successful approaches. Finally, Sect. 5 concludes the paper and indicates directions for our future research on this topic.

2 Knowledge Pit Data Challenge Platform

Knowledge Pit[1] is a web platform created to support organization of data mining challenges. On the one hand, this platform is appealing to members of the machine learning community for whom competitive challenges can be a source of new interesting research topics. Solving real-life complex problems can also be an attractive addition to academic courses for students who are interested in practical data mining. On the other hand, setting up a publicly available competition can be seen as a form of outsourcing the task to the community. This can be highly beneficial to the organizers who define the challenge, since it is an inexpensive way to solve the problem which they are investigating. Moreover, an open data mining competition can bring together domain experts and data analysts, which in a longer perspective may leverage a cooperation between the industry and academic researchers.

The Knowledge Pit platform is designed in a modular way, on top of an open-source e-learning platform *Moodle.org* and as such, it follows the best practices of a software development. The current modules of the platform include user accounts management system, competition management subsystems, time and calendar functionalities, communications features (i.e. forums and messaging subsystems), and a flexible interface for connecting automated evaluation services prepared to assess contestants' submissions.

Figure 1 shows an architecture schema of the Knowledge Pit platform. Its two main parts are the platform's engine located at a dedicated server and the evaluation subsystems. Currently, Knowledge Pit is hosted on a server belonging to Polish Information Processing Society[2].

The two main parts of the platform are the platform's engine and the evaluation subsystems. The first one provides interfaces for defining and maintaining of data challenges, management of user's profiles, submissions and private files, maintaining *Leaderboards* and the internal messaging systems (competition forums, chats, email and notification sending services, etc.). It is based on a very popular solution stack, i.e. Apache, MySQL and PHP. Together they constitute a bridge between the platform and different groups of users (guests, participants of competitions, moderators and organizers of particular challenges, managers and administrators of the system).

The second part of the platform is responsible for assessment of solutions submitted by participants of particular competitions. Due to a flexible communication mechanism, this service may be distributed among several independent

[1] https://knowledgepit.fedcsis.org.
[2] http://pti.org.pl/English-Version.

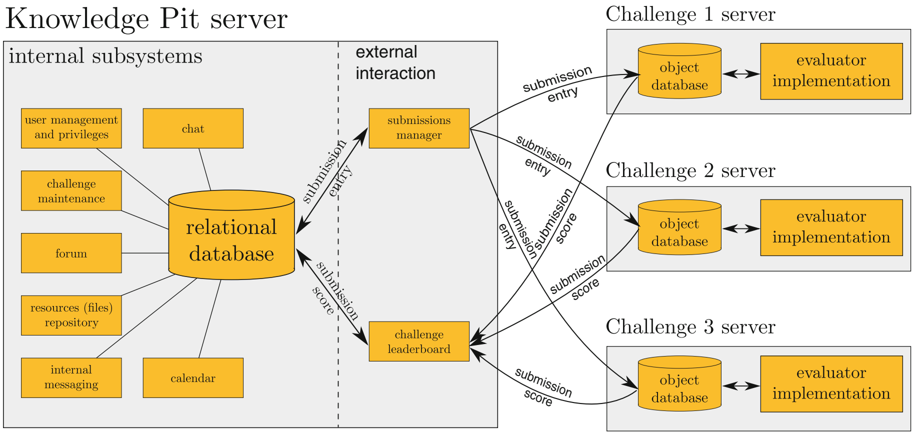

Fig. 1. A system architecture of the Knowledge Pit web platform.

workstations, which guarantees the scalability of the evaluation process. Since evaluating submissions for some competitions may require a lot of resources (e.g. memory, CPU time, disc I/O or database connections), this is a very important aspect of system's architecture. For example, the assessment of a single submission to AAIA'14 Data Mining Competition[3] required constructing several Naive Bayes classification models for a data table consisting of 50,000 objects and testing their performance on a different table with 50,000 objects described by 11,852 conditional attributes [5]. In that case, distribution of the required computations allowed for nearly real-time evaluation, even during the most busy moments when the rapid system load peak was observed.

Another advantage of separating the evaluation subsystems from the platform's engine is that it may be implemented in any suitable programming language, as a script or a standalone compiled application that can use any external libraries. In this way, the responsibility for preparation of a suitable evaluation procedure can be delegated to organizers of individual competitions. In such a case, the only requirement for the implementation of the evaluator is that it should maintain a correct protocol of information exchange with the platform's engine. The proposed flow of responsibilities releases the Knowledge Pit platform from the specific requirements and responsibilities which could not be processed in a generic way. It also gives competition organizers a very flexible method of expressing their data mining task in a form of a fully customizable evaluation procedures. For instance, the solutions submitted to IJCRS'15 Data Challenge[4], that is described in this paper, were evaluated using a script written in R language [6], which was running on an external server.

[3] https://knowledgepit.fedcsis.org/contest/view.php?id=83.
[4] https://knowledgepit.fedcsis.org/contest/view.php?id=109.

3 Scope of the IJCRS'15 Data Challenge

One of the main goals that we wanted to achieve was to increase the involvement of the community in devising efficient methods for assessing natural hazards in coal mines. As was explained in Sect. 1, a good solution to this issue could not only save human lives but it would also help to minimize financial losses caused by suspensions in the mining process. The task in our competition was to come up with a prediction model which could be efficiently integrated with a safety monitoring system of a coal mine, in order facilitate foreseeing warning levels of methane concentrations at three crucial methane meters located in vulnerable points of a longwall. In this section we describe the data which we provided for the competition and discuss the chosen evaluation criteria.

3.1 The Competition Data

The data used in the competition came from an active Polish coal mine. They correspond to a mining period between March 2, 2014 and June 16, 2014. The main data file for the competition consisted of multivariate time series corresponding to readings of sensors used for monitoring the conditions at the longwall. It was provided in a tabular format. In total, in the training data set there were series from $51,700$ time periods, each 10 min long, with measurements taken every second (600 values in a single series for every sensor). The values for each time period were stored in a different row of the data file. Each of the rows contained readings from 28 different sensors thus, in total, the data consisted of $16,800$ numerical attributes. The time periods in the training data were overlapping and given in a chronological order.

Data labels indicated whether a warning threshold had been reached in a period between three and six minutes after the end of the training period, for three methane meters: $MM263$, $MM264$ and $MM256$. If a given row corresponds to a period between t_{-599} and t_0, then the label for a methane meter MM in this row is "warning" if and only if $\max\left(MM(t_{181}), ..., MM(t_{360})\right) \geq 1.0$. The labels for the training data were provided in separate files. The test data file was in the same format as the training data set, however, the labels for the test series were not available for participants during the course of the competition. It is important to note that time periods in the test data did not overlap and they were given in a random order. This temporal disjunction between the training and test data makes the common assumption regarding i.i.d. data unfulfilled [7] and constitutes the biggest difficulty in the considered task.

In the data files provided for participants of the competition we also included a mining scheme utilized in the mine (Fig. 2), which explains the placement of all sensors used for monitoring the mining process. This file, combined with the provided description of sensors, constituted an important source of domain knowledge which was necessary to understand the dependencies between readings of different sensors. The cutter loader moves along the longwall between the sensors $MM262$ and $MM264$. The bigger are the currents consumed by the cutter loader, the more efficient is its mining work, which in theory results

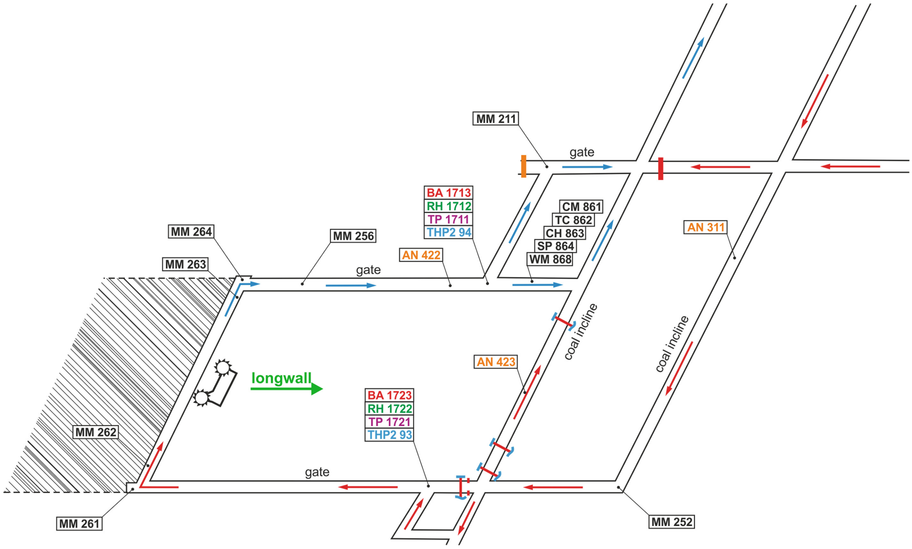

Fig. 2. A scheme of the mining process that generated the competition data.

in more methane emitted to the air. The arrows on the provided scheme show the directions of the air flow (along with methane flow) in the corridors. If the methane concentration measured by any of the sensors reaches the alarm level, the cutter loader is switched off automatically. However, if we were able to predict ahead the warning methane concentrations, we could reduce the speed of the cutter loader and give the methane more time to spread out – before the necessity of switching off the whole production line.

All the data for the competition were provided by Research and Development Centre EMAG[5] which was also the main sponsor of the competition.

3.2 Evaluation of Submissions

The evaluation procedure in the competition was two-fold. During the course of the contest an on-line evaluation system was providing constant feedback for the participants in a form of a publicly available Leaderboard — a dynamic ranking of participants' best results. However, the scores displayed on the Leaderboard were only preliminary assessments of the solutions' quality. They were computed using only 20 % of the available test data. The second evaluation round was conducted after completion of the competition. It was available only for those teams which had provided a short description of their solution in a form of a competition report. The final evaluation was carried out independently from the preliminary one, using the remaining part of the test data set.

[5] http://www.ibemag.pl/index.php?l=ang.

Table 1. Proportions of examples from the "warning" class for each of the target sensors.

Data set\sensor ID	$MM263$	$MM264$	$MM256$
Training set	0.009	0.026	0.025
Preliminary test set	0.007	0.023	0.025
Final test set	0.007	0.025	0.027

The quality criterion chosen for the evaluation of submissions was based on a well known Area Under ROC Curve (AUC) measure. This measure was selected due to the sparsity of positive examples from the "warning" decision class for each of the considered target sensors. Table 1 shows proportions of the cases from the "warning" class for the methane meters $MM263$, $MM264$ and $MM256$ in the training and test parts of the data.

A correctly formatted submission consisted of a single text file in the *csv* format. Consecutive rows of this file corresponded to cases from the test set. They ought to contain three real numbers indicating likelihoods of the label "warning" for the three target methane meters. The values did not have to be in a particular range, however, higher numerical values should have indicated higher chances of the label "warning". For each of the submitted file, AUC values were computed independently for each of the target sensors. The score received by a submission was a simple arithmetical mean of the obtained AUC values.

4 Results of the Competition

IJCRS'15 Data Challenge attracted many skilled data mining practitioners who managed to submit a variety of interesting solutions. There were 90 registered teams with members from 18 different countries. In total they submitted $1,676$ solutions. Additionally, 40 teams provided a brief report describing their approach. These reports turned out to be a valuable source of knowledge regarding the state-of-the-art in the predictive analysis of time series data.

In order to stimulate the competitiveness among participants and to give a good reference point to their results we provided a simple baseline solution. It was based on a technique that derives from the rough set theory [8]. The likelihood predictions for the target decision classes were made using an ensemble of 10 sets of decision rules. The sets were obtained by applying *LEM2* rule induction algorithm [9] to the training data discretized using the greedy discernibility-based heuristic [10,11]. The whole prediction model, including preprocessing of data, discretization, computation of decision rules and prediction of the likelihoods, was implemented in R System using the *RoughSets* package [12].

The top-ranked participants exceeded the score of the baseline solution by nearly 6 percentage points. Table 2 shows names of teams that achieved the best results in the competition.

Table 2. The final and preliminary results of the top-ranked teams.

Rank	Team name	Preliminary	Final score
1	Zagorecki	0.9666259	0.95926715
2	Marcb	0.94607103	0.94392893
3	Dymitrruta	0.93371596	0.94369948
4	Moomean	0.92862237	0.94280921
5	Trzewior	0.94691336	0.94134706
6	Kkurach_kp7	0.96853101	0.94002446
...	...	...	...
36	Baseline	0.8930246	0.90044912

The submitted solutions proved to be a valuable source of knowledge. They not only provided an insightful view on the state-of-the-art in multidimensional time series analysis but also contained inspiring ideas, designed specifically for the considered problem. The most interesting of these ideas are described by their authors in separate papers submitted to the competition track of IJCRS'15.

Investigation of the results revealed big differences between scores obtained by the participants on the preliminary and test sets. Since every team could submit more than one solution, it seems that some participants tried fine-tune their algorithms using the feedback from the preliminary test data. In theory, this approach could lead to a strong overfitting to that small set, which could explain the differences in results. However, a statistical test did not reveal a significant divergence between the differences in the preliminary and final results of the three top-ranked and the lower-ranked participants. The average difference was lower for the top-ranked teams but the p-value of the test was 0.3573.

We decided to compare submission histories of some of teams with the most interesting results. Figure 3 shows scores (both preliminary and final) of top 10 and last 10 solutions submitted by the winning team and a team from the third ten of the ranking. This team had a similar number of submissions as the winner and its best score (on the final test set) was also close to the maximum. It is important to notice that participants were not aware of final scores received by their submissions until the end of the competition and only one solution was taken into account for the final ranking. The final solution was either manually selected by participants or was automatically chosen based on its preliminary score. This comparison shows an interesting phenomenon – in a vast majority of cases the top-ranked teams achieved better results of the final test set than on the preliminary one, whereas for the lower-ranked teams there was an opposite tendency. Finding an explanation to this fact could be very important for the future development of predictive models for multivariate time series data.

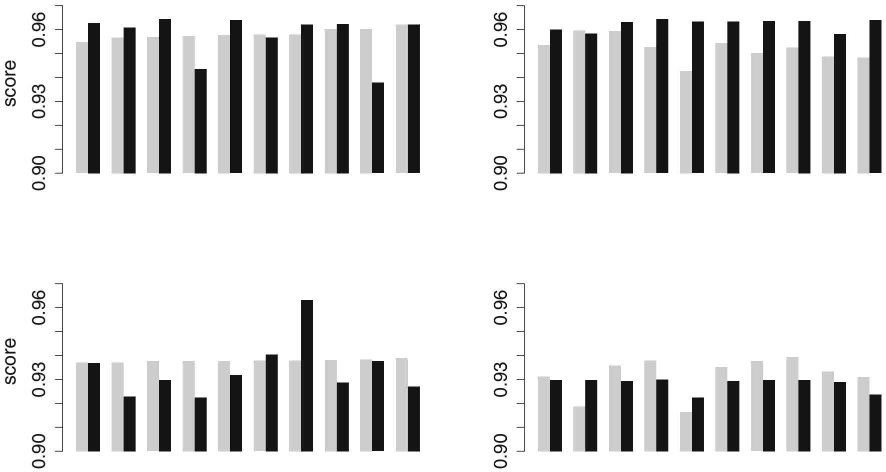

Fig. 3. The top left plot shows the preliminary (gray bars) and final scores (black bars) of top 10 solutions submitted by the team "zagorecki" (selected based on the preliminary results). The top plot on the right shows the scores of the 10 latest submissions from this team. For comparison we also present similar plots for a team placed in the third ten of the ranking.

5 Summary and the Future Research

IJCRS'15 Data Challenge was an on-line data mining competition hosted by the Knowledge Pit platform. It was held between April 13 and June 25, 2015. The task was to come up with a prediction model which could be applied to foresee warning levels of methane concentrations at a longwall of a Polish coal mine.

In this paper we explained our motivation for organizing this event and also provided details regarding the competition data sets and the task. We believe that our effort will result in stimulating research within the machine learning community in this very important field of study. We are planning to make available all the competition data, including labels for the test cases. This will make it possible to facilitate a collaboration between research teams that did not participate in the competition but are interested in this subject.

We are also planning to organize a continuation of IJCRS'15 Data Challenge. This time the data will be about a seismic activity and its impact on the safety in coal mines. The task will be to devise a classification model that can accurately predict energy of seismic shocks that will be perceived at longwalls of a mine. We hope that the success of IJCRS'15 Data Challenge will help us in attracting even more participants to the new competition.

References

1. Kozielski, M., Skowron, A., Wróbel, Ł., Sikora, M.: Regression rule learning for methane forecasting in coal mines. In: Kozielski, S., Mrozek, D., Kasprowski, P., Malysiak-Mrozek, B., Kostrzewa, D. (eds.) BDAS 2015, pp. 495–504. Springer, Cham (2015)
2. Krasuski, A., Jankowski, A., Skowron, A., Ślęzak, D.: From sensory data to decision making: a perspective on supporting a fire commander. In: Proceedings of WI-IAT 2013 Workshops, pp. 229–236. IEEE (2013)
3. Grzegorowski, M., Stawicki, S.: Window-based feature extraction framework for multi-sensor data: a posture recognition case study. In Ganzha, M., Maciaszek, L.A., Paprzycki, M. (eds.) Proceedings of FedCSIS 2015. IEEE (2015)
4. Kabiesz, J., Sikora, B., Sikora, M., Wróbel, Ł.: Application of rule-based models for seismic hazard prediction in coal mines. Acta Montanist. Slovaca **18**(4), 262–277 (2013)
5. Janusz, A., xc, A., Stawicki, S., Rosiak, M., Ślęzak, D., Nguyen, H.S.: Key risk factors for polish state fire service: a data mining competition at knowledge pit. In: Ganzha, M., Maciaszek, L.A., Paprzycki, M., (eds.) Proceedings of FedCSIS 2014, pp. 345–354. IEEE (2014)
6. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008)
7. Boullé, M.: Tagging fireworkers activities from body sensors under distribution drift. In: Ganzha, M., Maciaszek, L.A., Paprzycki, M. (eds.) Proceedings of FedCSIS 2015. IEEE (2015)
8. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Inf. Sci. **177**(1), 3–27 (2007)
9. Grzymała-Busse, J.W.: A new version of the rule induction system LERS. fundamenta Informaticae **31**(1), 27–39 (1997)
10. Nguyen, H.S.: On efficient handling of continuous attributes in large data bases. Fundamenta Informaticae **48**(1), 61–81 (2001)
11. Janusz, A.: Algorithms for similarity relation learning from high dimensional data. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets XVII. LNCS, vol. 8375, pp. 174–292. Springer, Heidelberg (2014)
12. 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)