

Predicting Dangerous Seismic Events: AAIA'16 Data Mining Challenge

Andrzej Janusz*, Dominik Ślęzak*[‡]

*Institute of Informatics, University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
{janusza,slszak}@mimuw.edu.pl

[‡]Infobright Inc.

ul. Krzywickiego 34, lok. 219, 02-078 Warsaw, Poland

Marek Sikora^{†§}, Łukasz Wróbel^{†§}

[†]Institute of Informatics, Silesian University of Technology
ul. Akademicka 2A, 44-100 Gliwice, Poland
marek.sikora@polsl.pl

[§]Institute of Innovative Technologies EMAG
Leopolda 31, 40-189 Katowice, Poland

Abstract—This paper summarizes AAIA'16 Data Mining Challenge: Predicting Dangerous Seismic Events in Active Coal Mines which was held between October 5, 2015 and March 4, 2016 at the Knowledge Pit platform. It describes the scope and background of this competition and explains our research objectives which motivated the specific design of the competition rules. The paper also briefly overviews the results of this challenge, showing the way in which those results can help in solving practical problems related to the safety of miners working underground. In particular, our analysis focuses on applications of prediction models in order to facilitate the assessment of seismic hazards, in a situation when the exploration of a given working site has just started and there is very little historical data available.

Keywords—data mining competition; multivariate time series data; attribute engineering; cold start problem; hazards assessment;

I. INTRODUCTION

THE COAL MINING is one of the most important industries which according to a report by *IBISWorld* employs worldwide over 3.5M people [1]. The exploration of coal often requires working in hazardous conditions. Miners in an underground coal mine can face many threats, such as, e.g. methane explosions, rock-burst or seismic tremors. To provide protection for people working underground, systems for active monitoring of the coal extraction processes are typically used. One of their fundamental applications is to screen the seismic activity in order to minimize the risk of severe mining incidents. To facilitate this task, data exploration and decision support tools can be employed, e.g. for predicting seismic activity in the nearest future.

From a data processing point of view, a decision support system which could aid in active monitoring of the coal mining process requires efficient methods for handling continuous streams of data [2]. 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 [3]. 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 [4]. The system developed in the frame of DISESOR project 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, such as those related to increased seismic activity. The idea to popularize this topic among the data science community by organizing open data mining challenges originated within this project.

The competition described in this paper is the second one in the series. The first one – IJCRS'15 Data Challenge – was focused on the problem of active monitoring and prevention of dangerous methane outbreaks [5]. The task was to design an efficient classifier for multivariate time series data that is generated by various sensors placed in corridors of underground coal mines. The main difficulty in that task was related to the problem of, so called, concept drift [6] and the necessity of constructing robust representation of the available data [7]. This competition was hosted by the Knowledge Pit platform [8] which supports the organization of data mining competitions associated with data science-related conferences.

Following the success of our first competition, AAIA'16 Data Mining Challenge was also organized at Knowledge Pit. This time, however, the task was related to the problem of foreseeing periods of increased seismic activity, that may endanger miners working underground. The main motivation for organizing this challenge as an open on-line competition was the fact that such an approach allows to conveniently review and evaluate performance of the available state-of-the-art methods. It is also an objective way of verifying not only a viability of the predictive models but also whole analytic processes which include preprocessing, feature extraction, model construction and post processing of predictions (e.g. ensemble approaches). Additionally, a huge influence on the final shape of AAIA'16 Data Mining Challenge had our research interest in a severity of the cold start problem for prediction models. In the coal mining context, this problem appears in a situation when the exploration of a given working site has just started and there is very little historical data available that can be

utilized for a construction of the prediction model for the assessment of seismic hazards.

In the following Section II we reveal details regarding the organization of the data mining competition and then, in Section III, we describe its course and results, including a brief characteristic of the most interesting approaches among the submitted solutions. Next, in Section IV, we show how the competition results were used to conduct an analysis of the cold start problem in the prediction of seismic hazards. Finally, we conclude the paper in Section V by drawing our plans for a continuation of this study.

II. AAIA'16 DATA MINING CHALLENGE

AAIA'16 Data Mining Challenge: Predicting Dangerous Seismic Events in Active Coal Mines took place between October 5, 2015 and February 27, 2016. It was organized under auspices of 11th International Symposium on Advances in Artificial Intelligence and Applications (AAIA'16, <https://fedcsis.org/2016/aaia>) which is a part of the FedCSIS conference series.

The task in this competition was related to the assessment of safety conditions in underground coal mines with regard to a seismic activity and early detection of seismic hazards. In particular, the data set provided to participants was composed of readings from sensors, such as geophones, that monitor the seismic activity perceived at longwalls of different coal mines and measure energy released by, so called, seismic bumps. Each case in the data was described by a series of hourly aggregated sensor readings from a 24 hour period. The provided data also contained information regarding the intensity of recent mining activities at the corresponding working site, coupled by the latest assessments of the safety conditions made by mining experts. Moreover, to further enrich the available data, for each working site that occur in the data set some additional meta-data were made available, such as an identifier of the mine, an identifier of a region where the working site is located or a working site's height.

Participants of the competition were asked to design a prediction model that would be able to accurately detect periods of increased seismic activity. In particular, the target attribute in the provided data (the decision) indicated cases for which the total energy of seismic bumps observed in a following 8 hour period exceeded a warning level of $5 * 10^4$ Joules (i.e. energy released in the period starting after the last hour of aggregated readings describing the case and ending 8 hours later).

In total, the provided data was described by 541 main attributes and 6 additional features related to particular working sites. The competition's data correspond to over 5 years of readings, which to best of our knowledge makes this research the most comprehensive study related to this domain, conducted anywhere in the world.

The data set was divided into a training part which was made available along with the corresponding decision labels and a test part. The labels for the test set were hidden from participants. The division of cases between the training and test sets was made based on a time stamps. In particular, the training data set corresponded to a period between May 5,

2010 and March 6, 2014. It consisted of a total of 133,151 data rows, each corresponding to a different 24 hour period which were overlapping for consecutive cases. The test data covered the period between March 7, 2014 and June 24, 2015. Unlike for the training set, to facilitate an objective evaluation of solutions and prevent a common problem with, so called, data leakage [9], the test cases were not overlapping and provided in a random order. For this reason the test set used in the challenge was much smaller than the training data. It is important to notice, however, that even though it consisted of only 3,860 cases, the test set covered a period of nearly 16 months.

Table I shows some basic characteristics of data from each working site that occurs in the competition data. It is worth noticing that not all working site that are present in training data also appear in the test set and there are a few working sites that are present in the test data but not in the training set. Such a situation reflects a real-life problem when the exploration of coal shifts to a new site for which there is no data available. A similar issue can also be identified within other domains, e.g. recommender systems, and is commonly referred to as *the cold start problem* [10]. Noticeable is also the fact that the distribution of cases with the 'warning' decision label is quite uneven for different working sites.

TABLE I. BASIC CHARACTERISTICS FOR DATA OBTAINED FROM DIFFERENT WORKING SITES. THE FIRST COLUMN SHOWS WORKING SITES IDS, WHEREAS THE FOLLOWING COLUMNS PRESENT INFORMATION REGARDING INITIAL EXPERT ASSESSMENTS OF THE WORKING SITE'S SAFETY, NUMBER OF DATA SAMPLES IN THE TRAINING AND TEST SETS, AND THE PERCENTAGE OF CASES WITH THE 'WARNING' DECISION LABEL.

main working site ID	initial mining assessment	number of training cases	number of test cases	training warnings (percent)	test warnings (percent)
146	a	5591	98	0.0014	0.0000
149	b	4248	98	0.0718	0.0018
155	b	3839	98	0.1681	0.0094
171	a	0	49	0.0000	0.0000
264	b	20533	0	0.0039	0.0000
373	b	31236	0	0.0113	0.0000
437	b	11682	0	0.0041	0.0000
470	c	0	258	0.0000	0.0078
479	a	2488	35	0.0000	0.0000
490	a	0	160	0.0000	0.0500
508	a	0	58	0.0000	0.0172
541	b	6429	5	0.0087	0.0000
575	b	4891	253	0.0045	0.0012
583	b	3552	215	0.0021	0.0029
599	a	1196	363	0.0148	0.0289
607	b	2328	209	0.0000	0.0000
641	a	0	97	0.0000	0.0103
689	b	0	83	0.0000	0.1205
703	a	0	145	0.0000	0.0069
725	b	14777	330	0.0920	0.0021
765	a	4578	329	0.0000	0.0022
777	b	13437	330	0.0000	0.0009
793	b	2346	330	0.0000	0.0045
799	a	0	317	0.0000	0.0000
total	-	133151	3860	0.0226	0.0508

A. Evaluation of the uploaded solutions

Participants of the competition had to prepare their solutions in a form of predictions of a likelihood that a given record from the test set has the label 'warning' and send their solutions using the submission system of Knowledge Pit. Each of the competing teams could submit multiple solutions. Quality of the submissions was measured using Area Under the ROC Curve (AUC) [11], [12]. The submitted solutions were

evaluated on-line and the preliminary results were published on the competition Leaderboard. The preliminary score was computed on a subset of the test set, fixed for all participants. Size of this subset corresponded to approximately 25% of the test set and it was composed of data from four working sites with different characteristics. The final evaluation was conducted after completion of the competition using the remaining part of the test data.

Apart from submitting their predictions, each team was also obligated by competition rules to provide a brief report describing its approach. Only the final solutions from teams which sent a valid report could undergo the final evaluation and be published among the competition results. In this way we were able to collect a vast amount of information regarding the current state-of-the-art in predictive analysis of multivariate time series data and objectively verify different methods of preprocessing, feature extraction and post processing of the predictions (i.e. ensemble approaches [13]).

B. A course of a competition

Since one of the main objectives in organization of AAIA'16 Data Mining Challenge was to investigate the cold start problem in the domain of natural hazard detection, we designed this competition in an uncommon way. To gather comprehensive data about an impact of the size of available data on quality of predictions for a given working site, the training data set described above was divided into five separate parts and the course of the challenge was split into six phases. Table II shows some basic participation statistics related to each of the phase.

TABLE II. BASIC PARTICIPATION STATISTICS FOR EACH PHASE OF THE CHALLENGE. IN THE LAST PHASE ALL TRAINING DATA WAS MADE AVAILABLE TO ALL PARTICIPANTS, REGARDLESS OF THEIR ACTIVITY.

	training set size (cases)	number of submissions	best preliminary score	best final score
phase 1	79893	99	0.9296	0.9290
phase 2	93211	278	0.9412	0.9320
phase 3	106527	1377	0.9452	0.9405
phase 4	119839	363	0.9451	0.9375
phase 5	133151	513	0.9452	0.9379
phase 6	133151	505	0.9452	0.9439

After the start of the challenge only the first part of the training data was revealed to participants. The four consecutive parts were made available in approximately monthly intervals (each interval corresponded to a new competition phase), however, only active teams that submitted a required number of files with predictions could access the new data.

In the sixth phase, which lasted for the last two weeks of the competition, all training data parts were revealed to all participating teams regardless of their previous activity in the challenge. It was done to equalize winning chances for teams that decided to join the competition in its latest period.

III. OVERVIEW OF THE COMPETITION RESULTS

AAIA'16 Data Mining Challenge attracted many skilled data mining practitioners who managed to submit a variety of interesting solutions. In total, there were 203 registered teams with members from 31 different countries. The most of participating teams were from Poland (106), however, there

were also many teams from countries such as India (14), United Kingdom (12), USA (12), Canada (9) and France (5).

Among the registered teams, 106 were active, i.e. submitted at least one solution to the Leaderboard. In total they submitted 3,236 solutions of which 3,135 were correctly formatted and successfully passed the evaluation procedure. Additionally, 50 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 related to early detection of seismic hazards.

TABLE III. FINAL RESULTS AND NUMBER OF SUBMISSIONS FROM THE TOP RANKED TEAMS. THE LAST ROW SHOWS RESULTS OBTAINED SOLELY FROM ASSESSMENTS MADE BY MINING EXPERTS, WHICH WERE AVAILABLE IN THE DATA (ATTRIBUTES *latest_seismic_assessment* AND *latest_comprehensive_assessment*)

team name	rank	n of submission	final result
snm (organizers)	–	2	0.9396
tadeusz	1	31	0.9393
deepsense.io	2	111	0.9384
yata	3	54	0.9342
podludek	4	71	0.9336
jellyfish	5	1	0.9335
millcheck	6	80	0.9329
kkurach	7	32	0.9312
gabd	8	21	0.9299
basakesin	9	30	0.9297
rough	10	4	0.9269
...
experts	(18)	–	0.9196

Table III shows scores achieved by the top-ranked teams. It is worth to notice that the highest result in the final evaluation was obtained by a team involved in DISESOR project and organization of the challenge (team *snm*). Its solution was created using feature extraction methods developed for the purpose of the DISESOR system [7], combined with a rough set approach to reducing data dimensionality [14] and an ensemble learning approach. In order to construct their solution, authors were using only the data available to all participants, however, due to their organizational involvement, team *snm* was excluded from the final ranking. More details regarding this solution can be found in [15].

Among the ranked teams, the highest score was obtained by the team *tadeusz* which was also a subset of the second team in the ranking – *deepsense.io*. Their solution was also based on an ensemble technique. In their approach, authors carefully select a subset of the training data which they later use for constructing and validating the prediction models. Moreover, authors make a significant effort to develop a procedure for an unbiased performance evaluation for tuning parameters of their models and the resulting ensembles. The whole approach is comprehensively described in [16].

In general, the overview of the most successful approaches used by participants suggests that the key steps to achieving good results in this task were:

- 1) Extracting relevant features (computing a new data representation) that aggregate time series data and are robust with regard to a concept drift.
- 2) Designing an appropriate evaluation procedure for testing performance of used prediction models and tuning their parameters.
- 3) Using an ensemble learning techniques for blending predictions of simpler models.

Moreover, the results clearly showed that the proposed task proved to be a challenging one for the most of participants. From the 106 teams that submitted at least one solution only 18 were able to outperform in the final evaluation a simple scoring model that was based on safety assessments made by mining experts. These evaluations were available in the data as two attributes, namely *latest_seismic_assessment* and *latest_comprehensive_assessment*. Even though these features could take only four ordinal values ($a < b < c < d$), a simple logistic regression model that utilizes those two features achieves AUC score of 0.9196 on the final evaluation data (0.9028 on the preliminary test set).

The most likely reason for the weaker results of a large share of participants is over-tuning of their models to the preliminary evaluation set. In a case of many teams, preliminary results were much higher than the final scores – the biggest difference was as high as 0.174 (over 17 percentage points). Noticeable is also the fact that in the preliminary evaluation 64 teams obtained a score which was higher than the score of the model based solely on the assessments of experts.

IV. ANALYSIS OF THE COLD START PROBLEM

The cold start problem is an important practical issue that is related to real-world applications of many decision support systems. In the case of coal mining, it typically appears when a system for monitoring natural hazards becomes operations for new, previously unexplored longwalls. One of our research objectives motivating the organization of AAIA'16 Data Mining Challenge was to investigate severity of this problem in the context of systems for early detection of periods of increased seismic activity.

For this reason the competition was divided into phases, as it was described in Section II (see Table II for details regarding availability of training data in consecutive phases). Since in each phase a new subset of training data was made available to active participants, we were able to verify the impact of this additional information by examining quality of solutions submitted in consecutive phases. Moreover, thanks to the competition rules that encouraged active participation, we received a large number of diverse solutions for analysis.

Figure 1 presents a distribution of evaluation scores obtained by submissions during the course of the competition. For this analysis we only used valid solutions with a reasonable quality (we disregarded 'random' submissions and those which obtained the preliminary score lower than 0.65). On that plot, black vertical lines denote dates on which additional parts of the training data set were released. Each solution on that plot is marked with a blue and red bar whose height corresponds to the obtained evaluation score. The level of red color in a bar indicates the final score, whereas the level of blue color marks the preliminary evaluation score.

A detailed analysis of the distribution of scores in time reveals some interesting observations. Firstly, in consecutive phases there is a quite conspicuous decrease in differences between the preliminary and final scores. In fact, in early phases of the competition preliminary scores tended to be much higher than the final ones, whereas in the last phase the trend was opposite. In order to confirm the statistical significance of this observation, we used a Wilcoxon rank sum

test of preliminary and final scores in consecutive phases. The test confirmed that average differences in phases 1, 2 and 3 and significantly higher (p -value $\ll 0.01$) than for the phases 4, 5 and 6. Interestingly, in the last phase the differences become negative (final scores are usually higher than the preliminary ones). This phenomenon can be explained by the fact that in the last few days of every data mining competition participants tend to focus on maximizing their score by blending their previous solutions. For this reason we will exclude the last phase from our further analysis of the cold start problem. Table IV shows mean and standard deviation of evaluation scores for each of the competition phases.

TABLE IV. MEAN AND STANDARD DEVIATION OF SCORES IN EACH OF THE COMPETITION PHASES. THE LAST COLUMN GIVES MEAN DIFFERENCES BETWEEN THE PRELIMINARY AND FINAL SCORES.

phase	prelim. mean	prelim. sd	final mean	final sd	mean diff.
phase 1	0.8590	0.0579	0.8251	0.0672	0.0339
phase 2	0.9059	0.0420	0.8851	0.0587	0.0207
phase 3	0.8683	0.0693	0.8307	0.1058	0.0376
phase 4	0.8868	0.0669	0.8772	0.0831	0.0096
phase 5	0.8943	0.0553	0.8857	0.0625	0.0086
phase 6	0.8820	0.0667	0.8942	0.0696	-0.0122

Another interesting observation related to analysis of the results shown on Figure 1 and displayed in Table IV is that the use of additional training data has a diminishing impact on performance of prediction models. For instance, if we compare average results from the second phase to results from the fourth or fifth phase, we see that the difference is minimal, even though in these phases we received a comparable number of submissions and the available training set data in, e.g. *phase 5*, was by nearly 43% larger than in *phase 2*. This was even less expected due to the fact that the data available in *phase 2* contained information about only 9 out of 21 main working sites present in the test data (these sites corresponded to $\approx 45\%$ of the test set), whereas in *phase 5* this number was much higher (13 out of 21 sites; $\approx 70\%$ of the test set).

To confirm the second observation, in each phase we analyzed the solutions with highest preliminary scores from teams that obtained scores higher than 0.85 – results of such teams better reflect performance of the state-of-the-art models. Figure 2 visualizes basic statistics (min,max,quantiles and mean values) for the preliminary and final evaluations of those submissions.

Conspicuous is the lack of significant differences in the best preliminary evaluation results in consecutive phases. The average final scores slightly increase from phase to phase, however, when we checked the statistical significance of the changes it turned out that a significant difference (p-value lower than 0.01) is only between results from the fifth and sixth phases. For other consecutive phases the p-value of Wilcoxon test was always higher than 0.175.

The above observations allow to formulate a hypothesis that having a sufficiently large data set it is possible to construct efficient prediction models for assessment of seismic hazards. The created models can outperform the currently used expert methods even for completely new working sites, as long as these sites have comparable geophysical properties and the same methodology is used for collecting new data. In order to verify this claim we decided to thoroughly investigate performance of top-ranked solutions submitted in each phase,

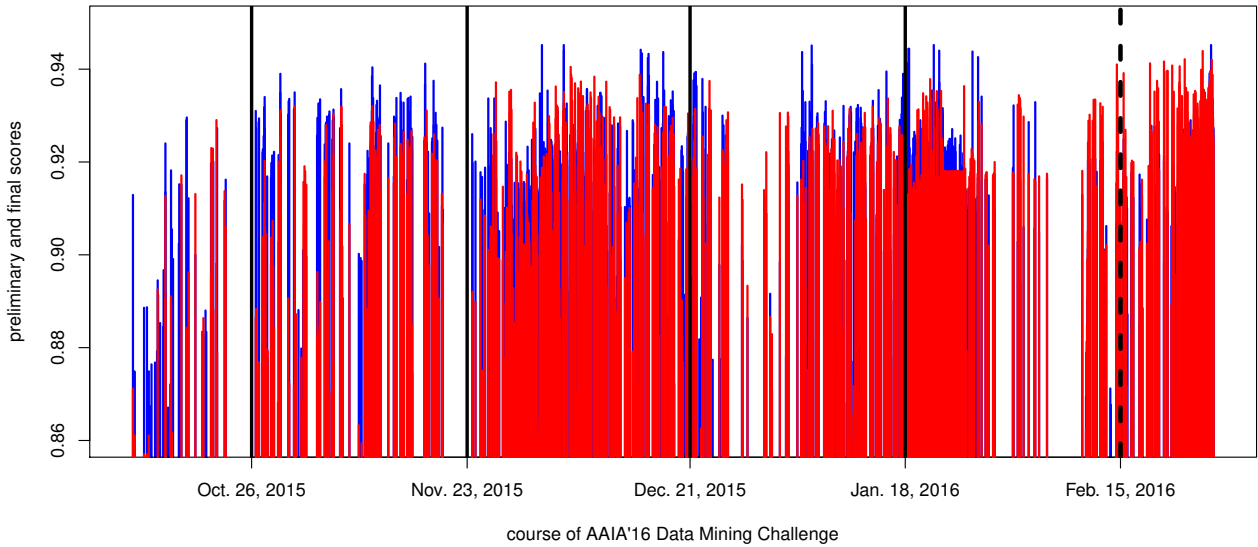


Fig. 1. Distribution of preliminary and final scores during the course of the competition. Blue bars show preliminary scores, whereas the corresponding red bars display final scores. The vertical black lines mark the dates which separate consecutive phases of the competition.

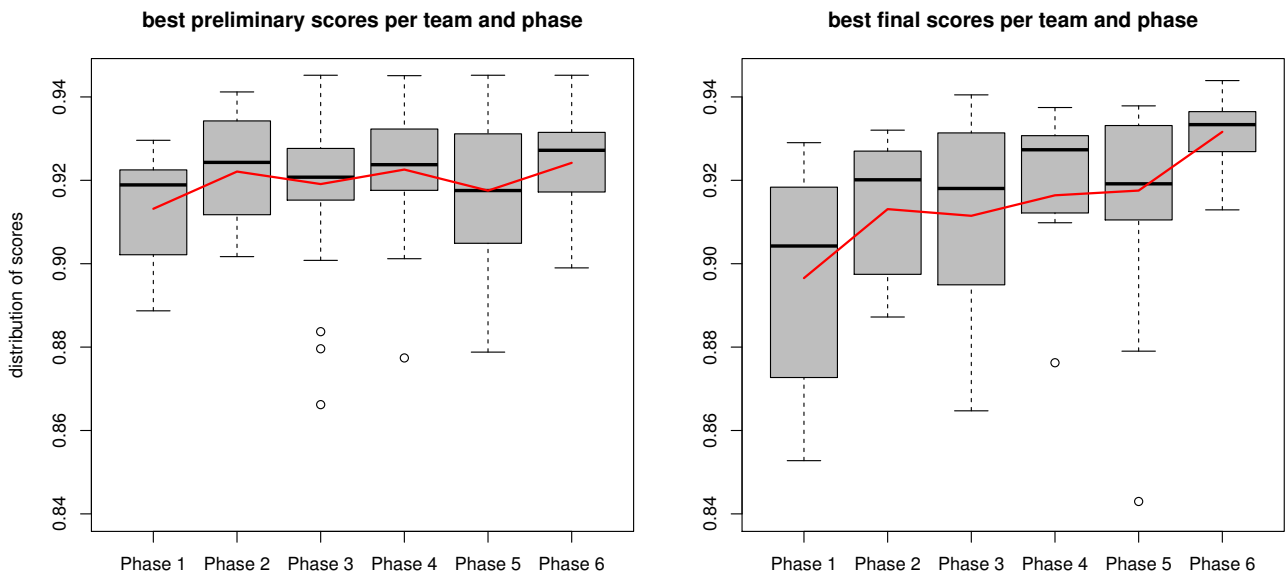


Fig. 2. Distribution of the best preliminary and final scores per team and competition phase. The red lines correspond to average values for given phases.

with regard to individual working sites.

For the purpose of this analysis we disregarded working sites for which there was no examples with the 'warning' label in the test set. The reason for that was the inability to compute values of AUC on such data subsets. In this way, for the remaining part of our analysis there were 15 working sites left, which corresponded to $\approx 81.5\%$ of the test data.

From solutions submitted in each competition phase, we have chosen 6 with the scores in top 10% for a given phase. During the selection process we considered only solutions

uploaded by teams actively participating in the competition, which fulfilled the criteria for obtaining all additional training data. Table V shows their average AUC values with respect to individual working sites. Additionally, the last two rows of the table give average values of AUC for working sites that are present in the training set and for those which are unavailable in the training data, respectively. Finally, the last column of Table V shows AUC values obtained for individual working sites using only the assessments made by experts.

For the most of working sites there is a statistically

TABLE V. AVERAGE SCORES OF TOP SOLUTIONS FOR INDIVIDUAL WORKING SITES, IN DIFFERENT PHASES OF THE COMPETITION. EVALUATIONS OF EXPERT ASSESSMENTS IS GIVEN FOR A COMPARISON IN THE LAST COLUMN. ADDITIONALLY, THE LAST TWO ROWS DISPLAY AGGREGATED VALUES (AVERAGES) FOR WORKING SITES WITH DATA IN THE TRAINING SET AND WITHOUT ANY AVAILABLE TRAINING DATA.

working site ID	phase 1	phase 2	phase 3	phase 4	phase 5	phase 6	expert assessments
149	0.8984	0.9056	0.8523	0.9062	0.8766	0.9005	0.9306
155	0.6578	0.7328	0.7492	0.7393	0.7242	0.7487	0.6845
470	0.9749	0.9922	0.9876	0.9935	0.9922	0.9964	0.9707
490	0.8013	0.8122	0.8340	0.8021	0.7892	0.8289	0.8109
508	0.9825	0.9971	1.0000	0.9942	0.9854	1.0000	1.0000
575	0.9348	0.9845	0.9859	0.9826	0.9820	0.9825	0.9723
583	0.9000	0.9419	0.9363	0.9388	0.9370	0.9401	0.9280
599	0.8391	0.8585	0.8678	0.8445	0.8670	0.8710	0.8020
641	0.9809	0.9983	1.0000	0.9965	1.0000	1.0000	1.0000
689	0.7723	0.8812	0.8523	0.8685	0.8582	0.8938	0.8884
703	0.9346	0.9792	0.9826	0.9699	0.9873	0.9722	0.9722
725	0.8968	0.9188	0.9251	0.9151	0.9176	0.9099	0.8955
765	0.7989	0.7911	0.7367	0.7608	0.7423	0.7808	0.7587
777	0.9118	0.9354	0.9242	0.9252	0.9175	0.9408	0.9444
793	0.9499	0.9545	0.9585	0.9538	0.9361	0.9468	0.8868
avail. in training	0.8653	0.8915	0.8818	0.8852	0.8778	0.8912	0.8670
unavail. in training	0.9077	0.9433	0.9428	0.9374	0.9354	0.9486	0.9404

significant improvement (tested using t-test with a confidence level of 0.05) of results from the later competition phases in comparison to the first phase. However, in nearly all cases the improvement between the second and later phases becomes marginal (one exception is the working site with ID 599). Interestingly, there are event sites (e.g. 689, 777) for which there is a noticeable drop in average quality of solutions between the second phase and phases 3, 4 and 5. Interesting is also the fact that the top solutions obtained consistently higher scores for working sites that were not present in the training data. Explanation of this fact require further analysis.

A comparison of the selected solutions to predictions that were based solely on assessments made by experts revealed that more complex models were able to quickly attain significantly higher scores for working sites with available training data. In the case of the remaining working sites the advantage of complex prediction models was not that clear. The average results for selected models in phase 6 were only slightly higher, however, for a part of investigated solutions the difference was much more favorable than for others.

V. CONCLUSIONS

In this paper we summarized AAIA'16 Data Mining Challenge: Predicting Dangerous Seismic Events in Active Coal Mines which was held at Knowledge Pit platform in association with 11th International Symposium on Advances in Artificial Intelligence and Applications (AAIA'16). We explained research goals that motivated us to organize this competition. We also explained the task in the challenge and briefly described its course. Finally, we showed a detailed analysis of competition's results with an emphasis on the cold start problem.

The conducted analysis revealed several interesting findings regarding the influence of additional training data on performance of prediction models for assessment of seismic hazards. It showed that in order to train prediction methods that aim to work well for a wide range of locations, it is sufficient to provide training data for only several different

working sites. Adding more data may have a minimal impact on prediction quality but it definitely helps in computing more reliable estimations of expected prediction performance, as well as in avoiding over-fitting of models to the training data.

Moreover, our analysis confirmed usefulness of the expert methods for assessment of natural hazards. Not only these assessments were able to robustly predict the seismic activity (they outperformed solutions of more than 80% of teams participating in the competition), but also they could be successfully applied to completely new working sites, without a need for using additional training data and complex algorithms.

This observation allows to formulate a general strategy for dealing with the cold start problem: for new working sites start predicting seismic hazards using the expert methods and concurrently gather data for training a more sophisticated prediction algorithm. Initiate your model using data from other working sites and then adjust it using the newly obtained data. Periodically compare performance of your model to results of the expert assessments and switch to your predictions when they become more accurate.

There are still several unanswered questions and research problems that we plan to investigate in our future work. For instance, the competition setting does not allow to study performance of incremental learning methods which can be applied to this problem. We would also like to more thoroughly analyze severity of the concept drift problem which in this context can be related to temporal nature of the data, as well as to changes in characteristics of different working sites. Another important issue is related to a development of methods for identification of good data subsets for training a prediction model for a given working site. Such methods could be based, for instance, on a comparison of similarities between different sites and choosing the data from those with the most similar characteristics.

Finally, in order to guarantee practical applicability of models for the mining industry it is important that mining experts could easily interpret and explain their predictions. For this reason, interpretability of a prediction model may be as important as its performance. The development of efficient algorithms that yield interpretable results is also directly related to a problem of extracting informative, yet compact representation of the training data. These two issues indicate prominent research directions for our future work.

ACKNOWLEDGMENTS

This research was supported by the Polish National Centre for Research and Development (NCBiR) grant PBS2/B9/20/2013 in the frame of the Applied Research Programme.

REFERENCES

- [1] IBISWorld. (2016) Global coal mining: Market research report. [Online]. Available: <http://www.ibisworld.com/industry/global/global-coal-mining.html>
- [2] A. Bifet and R. Kirkby, "Data stream mining: a practical approach," The University of Waikato, Tech. Rep., Aug. 2009.
- [3] J. Kabiesz, B. Sikora, M. Sikora, and Ł. Wróbel, "Application of Rule-Based Models for Seismic Hazard Prediction in Coal Mines," *Acta Montanistica Slovaca*, vol. 18, no. 4, pp. 262–277, 2013.

- [4] M. Kozielski, M. Sikora, and Ł. Wróbel, "Disesor - decision support system for mining industry," in *Proceedings of FedCSIS 2015*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2015, pp. 67–74. [Online]. Available: <http://dx.doi.org/10.15439/2015F168>
- [5] A. Janusz, M. Sikora, Ł. Wróbel, S. Stawicki, M. Grzegorowski, P. Wojtas, and D. Ślęzak, "Mining Data from Coal Mines: IJCRS&Ż15 Data Challenge," in *Proceedings of RSFDGrC 2015*, ser. LNCS, Y. Yao, Q. Hu, H. Yu, and J. W. Grzymala-Busse, Eds., vol. 9437. Springer, 2015, pp. 429–438.
- [6] M. Boullé, "Tagging Fireworkers Activities from Body Sensors under Distribution Drift," in *Proceedings of FedCSIS 2015*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds. IEEE, 2015, pp. 389–396.
- [7] M. Grzegorowski and S. Stawicki, "Window-Based Feature Engineering for Prediction of Methane Threats in Coal Mines," in *Proceedings of RSFDGrC 2015*, ser. LNCS, Y. Yao, Q. Hu, H. Yu, and J. W. Grzymala-Busse, Eds., vol. 9437. Springer, 2015, pp. 452–463.
- [8] A. Janusz, A. Krasuski, S. Stawicki, M. Rosiak, D. Ślęzak, and H. S. Nguyen, "Key Risk Factors for Polish State Fire Service: A Data Mining Competition at Knowledge Pit," in *Proceedings of FedCSIS'2014*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds. IEEE, 2014, pp. 345–354.
- [9] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *TKDD*, vol. 6, no. 4, p. 15, 2012. [Online]. Available: <http://doi.acm.org/10.1145/2382577.2382579>
- [10] L. H. Son, "Dealing with the new user cold-start problem in recommender systems: A comparative review," *Information Systems*, vol. 58, pp. 87–104, 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0306437914001525>
- [11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.
- [12] T. M. Mitchell, *Machine Learning*, ser. McGraw Hill series in computer science. McGraw-Hill, 1997.
- [13] A. Janusz, "Combining multiple predictive models using genetic algorithms," *Intelligent Data Analysis*, vol. 16, no. 5, pp. 763–776, 2012. [Online]. Available: <http://dx.doi.org/10.3233/IDA-2012-0550>
- [14] A. Janusz and D. Ślęzak, "Computation of approximate reducts with dynamically adjusted approximation threshold," in *Proceedings of ISMIS 2015*, F. Esposito, O. Pivert, M. Hacid, Z. W. Ras, and S. Ferilli, Eds., vol. 9384. Springer, 2015, pp. 19–28.
- [15] M. Grzegorowski, "Massively Parallel Feature Extraction Framework Application in Predicting Dangerous Seismic Events," in *Proceedings of FedCSIS 2016*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds. IEEE, 2016, in print September 2016.
- [16] R. Bogucki, J. Lasek, J. K. Milczek, and M. Tadeusiak, "Early Warning System for Seismic Events in Coal Mines Using Machine Learning," in *Proceedings of FedCSIS 2016*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds. IEEE, 2016, in print September 2016.