Proceedings of the Federated Conference on Computer Science and Information Systems pp. 249–252
DOI: 10.15439/2016F188
ACSIS, Vol. 8. ISSN 2300-5963

Application of RapidMiner and R Environments to Dangerous Seismic Events Prediction

Katarzyna Dusza*, Dominik Korda*, Krzysztof Kozłowski*, Bartłomiej Szwej*,
Michał Kozielski†, Marcin Michalak*‡, Marek Sikora*‡, Łukasz Wróbel‡
*Institute of Informatics, Silesian University of Technology
ul. Akademicka 16, 44-100 Gliwice, Poland
Email: {katadus637,domikor355,bartszw537}@student.polsl.pl
Email: krzysztofkozlowski1993@gmail.com
Email: {Marcin.Michalak, Marek.Sikora}@polsl.pl
†Institute of Electronics, Silesian University of Technology
ul. Akademicka 16, 44-100 Gliwice, Poland
Email: Michal.Kozielski@polsl.pl
‡Institute of Innovative Technologies EMAG
ul. Leopolda 31, 40-189 Katowice, Poland
Email: {Marek.Sikora, Lukasz.Wrobel, Marcin.Michalak}@ibemag.pl

***Abstract*—Underground coal mining is a branch of an industry which safety of operation is very dependent on the natural hazards. A proper seismic event prediction is a significant aspect of building classification models from the real data, which can affect the coal mining safety increase. In this paper four models, built in a well known data mining environments, are presented. The obtained models, depending on a given implementation of popular methods, occurred comparable to the best results from the competition.**

I. Introduction

THOUGH the production of energy from renewable resources has been increasing recently, the mining is still an important part of the industry. There are many countries — even in Europe — which produce most (e.g., 83% in 2012, Poland) or almost half (e.g., 45% in 2012, Germany) of its energy from coal. This means that problems of coal mining — especially the underground coal mining — still have a global meaning.

One of the aspects of safe and efficient coal mining is prediction of seismic hazards. Safety refers to saving workers from the accidents and injuries while efficiency refers to unplanned stops of longwall systems. Analysis and proper prognosis of potentially dangerous methane concentration [1, 2, 3] and seismic events [4, 5, 6] should improve the safety and reduce the costs of underground coal mining.

This paper presents several solutions of a problem of predicting dangerous seismic events in hard coal mines. This classification problem was a goal of a AAIA'16 competition for data scientists. The paper is organized as follows: it starts from a short description of a competition, data provided to the competitors and the evaluation method that was applied to submitted models. Then, the four models and the way of their development are described. The paper ends with a comparison of the result quality delivered by both the presented approaches and the contest winner followed by a final conclusions.

This work was partially supported by Polish National Centre for Research and Development (NCBiR) grant PBS2/B9/20/2013 within Applied Research Programmes. The infrastructure was supported by "PL-LAB2020" project, contract POIG.02.03.01-00-104/13-00.

II. Competition Task

This paper describes solutions submitted to the AAIA'16 data mining competition which summary is presented in [7]. Data for this edition of the competition came from the hard coal mining industry and was provided by Research and Development Centre EMAG. The main goal of the analysis was a prediction of dangerous seismic events in coal mines. The following part of the paper presents more detailed description of the data provided for the contest and a method of model evaluation. More detailed information about the competition can be found in [8].

The objective of each competitor was to devise a reliable prediction model able to detect periods of increased seismic activity that endangers miners working underground in coal mines.

A. Data

The competition training file contained 79,893 records, each corresponding to 24 hours of measurements. Values stored in a single record could be divided into two parts. The first one consisted of an identifier of the main working site and 12 other characteristics related to the whole period of 24 hours described by the record. The second part consisted of hourly aggregated measurements that count the number of seismic bumps perceived at longwalls and measure their total energy, thus, for each characteristic it included 24 consecutive values.

There was a total number of 541 columns in the data (including the main working site id). There was also available a separate file with additional information about all main working sites included in the data (in the training and test parts).

Labels in the data indicated whether a total seismic energy perceived within 8 hours after the period covered by a data record exceeded the warning threshold (i.e. $5 \cdot 10^4$ Joules). The labels of the test series were hidden from participants. It is important to note that time periods in the test data did not overlap and they were given in a random order.

An additional impediment for competitors and their models was the fact that the data was unbalanced. 78,722 records belonged to a "normal" class while the rest of them (only 1,171) was labelled as "warning".

The goal for the competition participants was to predict likelihood of the label "warning" for the records from the test set. For the consecutive objects exactly one real number corresponding to the predicted likelihood should be placed. The values did not have to be in a particular range, however, higher numerical values should indicate a higher chance of the label "warning".

B. Evaluation

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. It corresponded to approximately 25% of the test data. The final evaluation was done after completion of the competition using the remaining part of the test data. Those results were also published on-line. The assessment of solutions was done using the Area Under the ROC Curve (AUC) measure.

III. Overview of Solutions

The presented solutions were developed by students at the Institute of Informatics, Silesian University of Technology. Participation in the competition was an additional and optional activity for the students of the Computational Intelligence and Data Analysis course. The best achievements in the competition was promoted by the exemption from the final exam.

During the university course the students learn, among others, R [9] and RapidMiner [10] environments. Therefore, these two environments were applied by them to solve the competition task. Among the solutions presented in this paper one was developed in RapidMiner and the other three were developed in R environment. Besides, two solutions implemented the Artificial Neural Network (ANN) model and the other two implemented Boosted Trees model. The details of the data preprocessing and the model parameters are presented in the following paragraphs.

A. Solutions based on the Artificial Neural Network model

The presented solution was defined in the RapidMiner environment, where the process was based on the *Neural Net* operator. The whole process is presented in the Fig. 1. The usage of a *Nominal to Numerical* operator in the process was planned as a constant mapping of the consecutive (increasing) levels of threats a, b, c, d to the increasing integer values $1, 2, 3, 4$. In this approach the set of independent variables was reduced and therefore, the following variables were taken into consideration:

- latest_seismic_assessment,
- latest_comprehensive_assessment,
- max_gactivity.24,
- max_genergy.24,
- total_number_of_bumps.24.

These variables were determined by trial and error, starting from the attributes that are highly correlated with the predicted variable.

The final model of the network had 8 neurons in a hidden layer and the following set of initial parameters of the *Neural Net* operator was chosen:

- training cycles: 700,
- learning rate 0.05,
- momentum: 0.2,
- decay: False,
- shuffle: True,
- normalize: True,
- error epsilon: 1.0E-5,
- use local random seed: True,
- local random seed: 1337

The final prediction quality of this model — submitted by Krzysztof Kozłowski as *unnamed* to the Knowledge Pit platform — expressed by means of AUC criterion was calculated as 0.9215.

The second solution based on the ANN model was developed in R environment. The H2O platform [11] which can be used in R environment was chosen as an implementation of neural network engine. One of the reasons of this implementation selection was the fact that it is very well documented and many helpful remarks on neural network parameter tuning are available [12].

Due to the data structure where 24 hour measurements were contained in each record, it was required to aggregate information from 24 columns representing consecutive hours of a day into a single one. The other attributes were selected on the basis of their correlation. From the results of experiments it occurred that due to normalization of the data the quality of results decreased. Thus, this processing method was abandoned.

The following set of initial parameters of artificial neural network was chosen:

- neuron activation function: $\tanh$ with dropout,
- number of neurons in a hidden layer: 5,
- input neurons droput: 0.1,
- hidden neurons droput: 0.3,
- classes balancing: turned off,
- maximal number of epochs: 300,

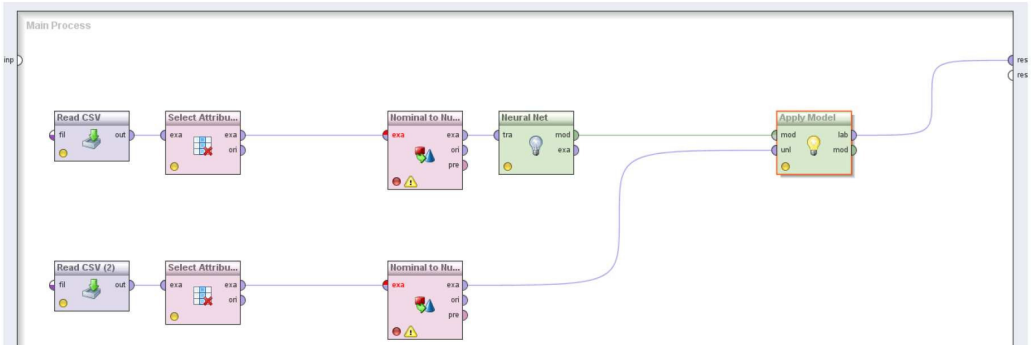

Fig. 1. Process of building the prediction model in RapidMiner.

The rest of the parameters was set to default values. The obtained neural network gave the following result expressed by means of AUC criterion: 0.9101.

Next, further tuning of parameters was performed what resulted in the following parameter values:

- neuron activation function: rectifier with dropout,
- number of neurons in a hidden layer: 4,
- input neurons droput: 0.2,
- hidden neurons droput: 0.3,
- classes balancing: turned off,
- maximal number of epochs: 250,
- $L1$ regularization: 10^{-5},
- adaptive speed of learning: turned off,
- number of training objects per iteration: 1000,
- number of testing objects: 80,
- maximum duty cycle fraction for scoring: 0.

This solution submitted by Dominik Korda and identified as *doxus* at the Knowledge Pit platform achieved the final prediction quality (AUC) value equal to 0.9225.

B. Solutions based on the Boosted Trees model

There are two approaches based on the Boosted Trees method presented in this section. Both of them were developed in R environment and both of them utilised a *caret* package [13]. The approaches differ due to the data processing stage.

Within the first approach the Kibana tool [14] was applied to visual analysis of the attributes. It enabled to notice an important association between the attribute *latest_seismic_assessment* and the decision attribute. Therefore, this attribute was included into the model in the first order.

Another important observation was connected with the *total_destressing_blasts_energy* attribute: for all objects that have a *warning* value of a decision, *total_destressing_blasts_energy* equals 0. Therefore, it was decided to introduce a new derived variable named *tdbeGTzero* (total destressing blasts energy greater than zero) defined in the following way:

$$\text{IF } total\_destressing\_blasts\_energy > 0$$

$$\text{THEN } tdbeGTzero = true$$

$$\text{ELSE } tdbeGTzero = false$$

The final set of the selected independent variables was as follows:

- sum_e3,
- sum_e4,
- sum_e5,
- sum_e6plus,
- highest_bump_energy,
- max_genergy,
- avg_genergy,
- tdbeGTzero,
- latest_seismic_assesment.

This solution, submitted by Bartłomiej Szwej and identified as *0bartek* at the competition platform, achieved the final prediction quality (AUC) value equal to 0.9238.

The set of independent attributes of the second approach was selected arbitrarily and it contained:

- latest_seismic_assessment,
- latest_seismoacoustic_assessment,
- latest_comprehensive_assessment,
- latest_hazards_assessment.

These attributes categorize the seismic activity into four levels (a, b, c, d), and a proper value is set by a domain expert working in a coal mine.

This solution, submitted by Katarzyna Dusza and identified as *kd* at the competition platform, achieved the final prediction quality (AUC) value equal to 0.9185.

IV. Results and Conclusions

Over one hundred (106) competitors accessed the challenge and 49 of them submitted their results. The quality of the top ten approaches and the solutions presented above are listed in Table I.

If we take into consideration all 49 results it can be stated that students' models placed higher than the median of all of them (25^{th} result was 0.91304342). This enables a positive assessment of these students' involvement to the contest. It is also worth to be noticed that some of them did not limit themselves only to tuning the method parameters but also tried to select and derive explainable attributes for the model.

TABLE I
SELECTED RESULTS FROM THE FINAL BOARD OF AAIA'16 DATA MINING CHALLENGE.

rank	participant	AUC
1	tadeusz	0.9393
2	deepsense.io	0.9384
3	yata	0.9342
4	podludek	0.9336
5	jellyfish	0.9336
6	millcheck	0.9329
7	kkurach	0.9312
8	gabd	0.9300
9	basakesin	0.9297
10	rough	0.9269
13	**0bartek**	0.9238
15	**doxus**	0.9225
17	**unnamed**	0.9215
18	**kd**	0.9185
49	researchlabs	0.6998

Additionally, it can be noticed that all the presented solutions were developed in a well known data mining environments. Therefore, these approaches are more general at the level of model creation, where only parameter tuning was performed. Besides, the presented solutions focused on a proper data pre-processing in order to select and derive the right independent variables.

Tuning of the parameters was performed in case of ANN-based approaches and the results presented in Tab. I show its positive impact. However, in case of the approaches based on the Boosted Trees model, where the parameters were identical and the results (see Tab. I) were significantly different, it is visible how important is the data processing phase of analysis.

Finally, from the university course leader perspective the involvement of the students into such data analysis competition is very promising. The students have the opportunity to operate on a real-life data and to compare the quality of their results with the other competitors. Therefore, it can be twofold interesting for them and hopefully it will increase their motivation to further studies.

REFERENCES

[1] M. Sikora and B. Sikora, "Improving prediction models applied in systems monitoring natural hazards and machinery," *International Journal of Applied Mathematics and Computer Science*, vol. 22, no. 2, pp. 477–491, 2012. doi: 10.2478/v10006-012-0036-3. [Online]. Available: http://dx.doi.org/10.2478/v10006-012-0036-3

[2] ——, "Rough natural hazards monitoring," in *Rough Sets: Selected Methods and Applications in Management and Engineering*. Springer, 2012, pp. 163–179. [Online]. Available: http://dx.doi.org/10.1007/978-1-4471-2760-4-10

[3] A. Zagorecki, "Application of sensor fusion and data mining for prediction of methane concentration in coal mines," *Mining — Informatics, Automation and Electrical Engineering*, vol. 524, no. 4, pp. 33–38, 2015.

[4] 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. 3, 2013.

[5] J. Kabiesz, "Effect of the form of data on the quality of mine tremors hazard forecasting using neural networks," *Geotechnical & Geological Engineering*, vol. 24, no. 5, pp. 1131–1147, 2006. doi: 10.1007/s10706-005-1136-8. [Online]. Available: http://dx.doi.org/10.1007/s10706-005-1136-8

[6] A. Leśniak and Z. Isakow, "Space-time clustering of seismic events and hazard assessment in the Zabrze-Bielszowice coal mine, Poland," *International Journal of Rock Mechanics and Mining Sciences*, vol. 46, no. 5, pp. 918–928, 2009. doi: 10.1016/j.ijrmms.2008.12.003. [Online]. Available: http://dx.doi.org/10.1016/j.ijrmms.2008.12.003

[7] A. Janusz and et al., "Predicting dangerous seismic events in active coal mines: Summary of AAIA'16 data mining competition at knowledge pit," *Proc of FedCSIS 2016*, vol. 00, no. 00, pp. 00–00, 2016.

[8] AAIA'16 data mining challenge: Predicting dangerous seismic events in active coal mines. [Online]. Available: https://knowledgepit.fedcsis.org/contest/view.php?id=112

[9] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org

[10] RapidMiner. Rapidminer. [Online]. Available: http://rapidminer.com

[11] H2O platform. [Online]. Available: www.h2o.ai

[12] The definitive performance tuning guide for h2o deep learning. [Online]. Available: http://blog.h2o.ai/2015/02/deep-learning-performance/

[13] The caret package. [Online]. Available: http://topepo.github.io/caret/index.html

[14] Kibana software. [Online]. Available: www.elastic.co/products/kibana