ADAM ZAGORECKI
Centre for Simulation and Analytics
Cranfield University
Defence Academy of the United Kingdom

Application of Sensor Fusion and Data Mining for Prediction of Methane Concentration in Coal Mines

In recent years we have experienced unprecedented increase of use of sensors in many industrial applications. Modern sensors are capable of not only generating large volumes of data but as well transmitting that data through network and storing it for further analysis. These enable to create systems capable of real-time data fusion in order to predict events of interest. The goal of this work is to predict methane concentration levels in coal mines using data fusion and data mining techniques. The paper describes an application of a generic method that can be applied to arbitrary set of multivariate time series data in order to perform classification or regression tasks. The solution presented here was developed within the framework of IJCRS'15 data mining competition and resulted in the winning model outperforming other solutions.

keywords: data mining, data science, prediction, methane outbursts.

1. INTRODUCTION

Because of the inherent dangers associated with underground coal mining, such as methane explosions or rock bursts, safety is one of paramount priorities for the coal mining industry. New technologies, such as networked sensors and predictive analytics can support the effort of providing safe environment for personnel underground. These technologies have potential not only providing reliable real-time environmental monitoring but as well providing advanced warning based on predictive numerical models [8,9]. In particular, data mining techniques allow for automated ways to create models derived from historical data rather than from expert knowledge. These are theoretically capable of identifying patterns that can be unknown to experts and that can lead to better predictions and often to gaining better insights into the domain.

The paper presents an application of a generic approach to classification of multivariate time series data to prediction of methane outbreaks. The presented approach was developed and evaluated in the context of the 2015 AAIA Data Mining Competition, where it led to the second highest score [10]. This paper concerns with applying the same fundamental approach to coal mine sensor data provided within the frame of another data mining competition: the IJCRS'15 Data Challenge: Mining Data from Coal Mines. The presented solution resulted with the winning entry. Even though the principles of the approach stay the same, in the process of developing models within the approach requires customization step which is specific to particular problem and characteristics of the data set. This is the customization step that can be of specific interest to the mining community – it sheds some light into characteristics of sensor readings that can provide early hints of dangerously increasing methane level in a several minute horizon.

The rest of the paper is composed as follows: in the next section a brief discussion on the technologies involved and the concept of data mining competitions will be presented. In the following section, the com-

petition task will be introduced with details of the sensors, available data and the evaluation. In the following section the proposed approach to classification of multivariate time series data will be discussed. Consequently each step in of the proposed approach will be discussed in more detail: feature engineering, and actual classification. The paper will finish with a short discussion.

2. BACKGROUND

During the recent decade affordable and reliable sensors capable of collecting large amounts of data have become popular in many applications including industrial, commercial and everyday life. One of most popular types of data collected by those sensors is time series data. This kind of data typically consists of sequences of measurements taken over time. With affordable sensors capable of transmitting data over network, multivariate time series data sets are becoming common. Examples can include vehicle or machinery monitoring, sensors from smart phones or sensor suites installed on human body. Because of the nature of time series, the collected measurements typically not directly exploitable – as the measurements consist of a large number of data points, it is often subject to noise, and require further analysis in order to identify or discover interesting patterns that can be exploited by users. It is well recognized that processing the raw measurements and transforming the data into knowledge useful for the users is a challenging and costly task. It is particularly true with multivariate time series data as time series are characterized by large volume and often need to undergo transformations (such as Fourier transforms, various filtering, etc.) to reveal potentially useful patterns. On the other hand, if generic methods for transforming multivariate time series data are developed, they can lead to rapid advances in utilization of sensor data in many areas. In this paper we present an application of a method for classification of multivariate time series data that was developed for a data mining competition involving motion sensors installed on human body and subsequently was successfully applied to different problem involving sensors installed in coal mines.

In the application presented in this paper the data consisted of measurements taken by various sensors installed on machinery operating in that coal mine and various locations in the coal mine. The sensors involved different types of measurements, mostly environmental such as humidity, temperature, air pressure, methane concentration, etc. and some related to the state of operating machinery such as cutter loader speed, direction and currents at different parts of machinery. The task was to predict if methane level exceeding certain thresholds would occur in next 3 to 6 minutes after the measurements were taken. This knowledge would potentially enable extra warning time before methane warning threshold is exceeded and can be used to take preventive actions.

In order to come up with a predictive model, a data mining competition was prepared within the frame of a scientific conference. Data mining competitions are often organized to encourage multiple scientists, students and hobbyists to tackle a given problem and identify the best method. They are similar to sport competitions, where the participants submit their solutions and the performance of their models is determined by the competition organizers to ensure independent and fair comparison. The results are published in a form of a leader board that is publicly announced. The participants can observe progress of the competition over time and make multiple submissions. A typical duration of a competition is measured in months, allowing for sufficient time to test different ideas and to develop more sophisticated and matured solutions. There are awards for the winners, which are typically monetary awards (typically order of several thousand dollars) and additional prizes such as free conference registration or computer hardware. The data mining competitions are highly regarded in academic community as fairly objective and realistic means to evaluate algorithms and ideas.

3. METHANE CONCENTRATION FORECASTING - IJCRS‘15 DATA MINING COMPETITION TASK

This paper presents a solution to the IJCRS‘15 data mining competition which was organized using the Knowledge Pit competition platform [6,7]. The objective of the competition was to gain insight into dependencies between cutter loader (mining machinery) performance and methane level measured by several sensors distributed in the coal mine.

The basic task of the competition was to create a numeric model to predict exceedance of threshold levels at three methane sensors in short future (3 to 6 minutes) based on sensors readings from multiple sensors.

The data used in the competition was collected from an active Polish coal mine during a the mining period between March 2, 2014 and June 16, 2014. The main data set for the competition consisted of

multivariate time series corresponding to sensors reading used for monitoring the environmental and mining machinery conditions at the longwall. A scheme showing mine layout including placement of environmental sensors was made available to the competition participants and is presented in Figure 1. Additionally, the organizers provided description of sensors in an effort to equip participants with understanding of the dependencies between readings of different sensors.

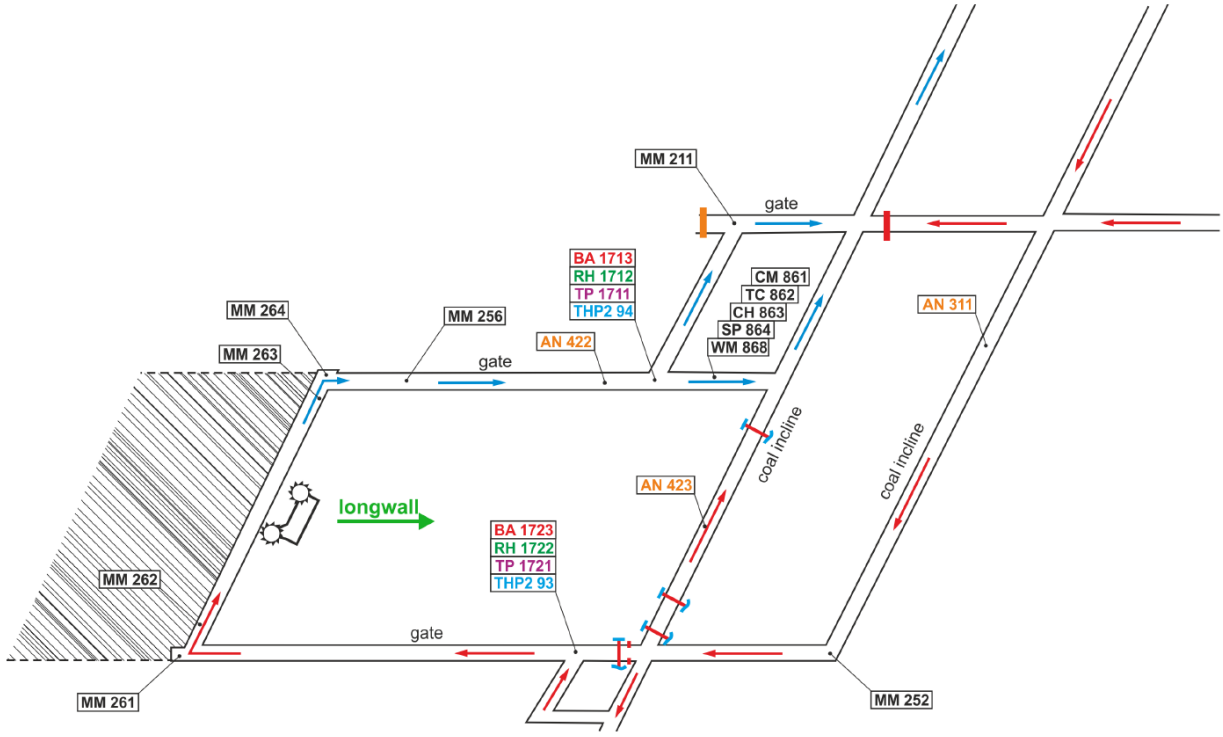

Fig. 1. A scheme of the mining process that generated the competition data

The cutter loader moved along the longwall between the sensors MM262 and MM264. The mining work of the cutter can be estimated by measuring electric currents and those measurements were available in the competition data. It is believed, that the more intensive the cutter's mining work is the more methane is emitted from the wall to the air. The air flow directions in the corridors were provided to the participants as shown in the scheme. If the methane concentration measured by any of the sensors reaches the alarm level, the cutter loader is shut down automatically, resulting with financial losses related to downtime.

That leads us to the practical benefit of predicting methane outbreaks – if one is able to reliably predict methane concentrations, then the cutter speeds can be adjusted to let the excessive methane concentrations to clear the area and avoid reaching the safety threshold and consequently leading to continuous operation.

All the data for the competition were provided by Institute of Innovative Technologies EMAG[1] which was also the main sponsor of the competition.

[1] www.ibemag.pl/en

3.1. Data

The data made available for this competition consisted solely of time series. It consisted of 51 700 records, which corresponded to a set of time series.

Each record consisted of 28 time series. Each time series was composed of exactly 600 data points and corresponded to a 10 minute worth measurements. The measurements included: anemometers, temperature sensors, methane sensors, barometers, humidity sensors, pressure and pressure difference sensors, current sensors (machinery), direction and speed of the cutter. In total, each record in the data set consisted of 16 800 numerical attributes.

The task was to predict methane level exceedance in the future at three methane sensors. The target variables were three binary attributes (threshold exceeded or not) for the three selected methane sensors. The three target attributes indicated whether a warning threshold for three methane sensors MM263, MM264 and MM256 had been reached in a period between three and six minutes after the end of the training period. 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 (MM(t_{181}), …, MM(t_{360})) ≥ 1,0 , which in practice meant that a methane warning level was exceeded within 3 to 6 minutes after the end of corresponding time series.

The original data was split into two sets: the training and test set. For the training set the target variables were provided. The task was to predict probability of the warning threshold exceedance for each of three target variables in the test set. The values of target variables for the test set were not available to participants and only the competition organizers had access to them.

It is worth noting that time series 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 [2] and constitutes the biggest difficulty in the considered task.

3.2. Evaluation

The evaluation of the results was performed using the Area Under the ROC Curve (AUC) measure concept [5]. It was possible, because the target variables were defined in form of probability of threshold exceedance for each of three variables. For each of three target variables the separate AUC score was first computed. Let us define the AUC score for the i^{th} target variable as AUC_i. The final score was the average of three individual scores:

$$AUC_{total} = \frac{1}{3}\sum_{i=1}^{3} AUC_i \qquad (1)$$

During the competition only preliminary score was available to competitors. The preliminary score was based on a subset of the final test set, and it corresponded to approximately 20% of the test data. The final evaluation was performed after completion of the competition.

4. SOLUTION OVERVIEW

In this section an overview of the solution to the competition task is presented. The method used was based on the method developed for the AAIA'15 data mining competition that is described in [10]. The key difference is that for IJCRS'15 competition it did not use feature selection step. It was decided not to use feature selection step as for the this competition feature selection resulted in inferior results and proved unnecessary. The basic steps in the used method are presented in Figure 2.

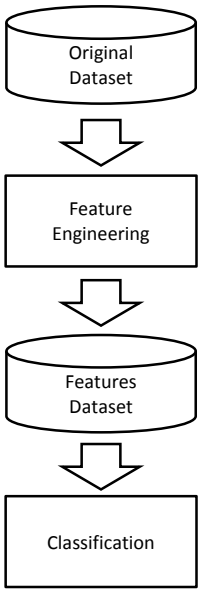

Fig. 2. The outline of the basic steps used during the competition

The first, and probably the most critical step was the feature engineering step. At this step the original data set was converted to a secondary data set that consisted of the features generated from the time series data. This step is discussed in detail in the Section 5.

For each of the three target variables a separate classifier that made a binary decision was decided to be created. This resulted with need to learn three separate classifiers (one for each decision variable). It is important to note, that no information was shared between three classifiers and all three of them were learned using the same features data set. The task called for assignment of probability (rather than hard decision) for presence of the threshold exceedance. As the basic classifier Random Forest was used. The benefit of using Random Forest is that it allowed to compute the probability of class assignment.

The original set of time series used was expanded for feature selection by creating additional time series that were derived from the original time series. The derived time series were generated from a pair of the original times series. Let us assume that x(t) and y(t) are two original time series, then the derived time series were generated if:

- Both x(t) and y(t) were methane sensors (their names started with M)
- Both x(t) and y(t) started with BA, RH, TP, and AN – all of them corresponded to particular type of environmental sensors: pressure, humidity, temperature, and wind speed.

– For each of the pair of signals x(t) and y(t) that met above conditions, two derived time series $d_1(t)$ and $d_2(t)$ were produced which were derived as follows:
  – $d_1(t) = x(t)-y(t)$ – a simple difference between corresponding measurements
  – $d_2(t) = (x(t)-y(t))/x(t)$ – a relative difference between corresponding measurements

This resulted with 52 derived time series.

5. FEATURE ENGINEERING

The next step was transformation of the data from time series form into a set of numerical values that summarize different aspects of the time series data.

The most basic features that can be derived from individual time series are simple statistics (e.g. mean, standard deviation), more complex features can be be derived from more than one time series (e.g. correlation coefficient between two time series). In the course of competition a lot of experimentation with different features was done. Weka software [3] feature selection algorithms were used to identify most informative features. As the result of this analysis a special emphasis on features related to maximal or minimal values, as those seemed to be most informative, at least according to feature selection algorithms, was put. I worth to be noticed, that the feature selection was used only to inform feature engineering– feature selection was not used to actually select features for classification –all generated features were used for classification task.

5.1. Generated Features

For each of the time series (either original or derived) the following features were extracted:

– the mean value
– the standard deviation
– the minimal value
– the maximal value
– the average of top 5 minimal values
– the average of top 5 maximal values
– the minimal value expressed in standard deviations from the mean
– the maximal value expressed in standard deviations from the mean
– the average of top 5 minimal values expressed in standard deviations from the mean
– the average of top 5 maximal values expressed in standard deviations from the mean
– the maximal difference between minimal and maximal values taken over non-decreasing sequences of measurements
– the maximal difference between maximal and minimal values taken over non-increasing sequences of measurements
– the maximal values (frequency and power) for the fast Fourier transform with ignoring first three frequencies
– the parameters for linear regression: slope, intercept, the mean square error, and the absolute value of slope
– the parameters for polynomial fitting (done only for parabolic fitting): a_0, a_1 and a_2
– the parameters for polynomial fitting taken over the first half of the signal (done only for parabolic fitting): a_0, a_1 and a_2
– the parameters for polynomial fitting taken over the second half of the signal (done only for parabolic fitting): a_0, a_1 and a_2

Each of the above features generated a single number that was used as an individual feature for further analysis. This produced a total of 2,214 features – 756 from the original time series and 1 458 from derived time series.

5.2. Correlations

Finally, it was decided to add correlation coefficients between time series. Additional parameters were derived from cross-correlations (those included autocorrelations) between selected pairs of signals:

– cross-correlations for the signal taken at t=0 and the same signal taken at t=0, 100, 200, and 300 using Pearsons' correlation coefficient
– cross-correlations for the signal taken at t=0 and the same signal taken at t=0, 100, 200, and 300 using Spearmans' correlation coefficient
– cross-correlations for the signal taken at t=0 and the same signal taken at t=0, 100, 200, and 300 using Kendalls' correlation coefficient

The pairs of signals x(t) and y(t) that were selected to compute correlation coefficients included:

– any methane sensors measurements MM taken pair-wise
– pairing signals starting with the same prefixes that were BA, RH, and AN
– pairs only if two signals had the same prefix – for example BA with BA, but not with any other

This effectively lead to include auto-correlation as it was allowed x(t)=y(t). The total number of features in the winning set was 4 914.

6. CLASSIFICATION

Random Forest [1] implemented in Weka software [3] were used as the basic classifier.

Also experimented with other classifiers such as Neural Networks, Logistic Regression, Support Vector Machines, and others were carried out, however the Random Forest seemed to perform consistently better. One of the challenges with applying Random Forest effectively is selection of optimal number of features used for each tree. In the case of competitions it is typically done by trial and error approach. Different numbers of features per tree were taken into consideration and for the particular feature set the numbers between 60 and 100 features seemed to work well. For the best score each of three Random Forest classifiers had 1 000 trees. The number of features for each tree was limited to 80.

7. CONCLUSIONS

In this paper an application of predictive analytics based on data mining techniques for predicting methane outbreaks in coal mines is described. The presented algorithm describes the winning solution for the IJCRS'15 Data Mining Competition. The solution is a customized approach to classification of multivariate time series developed for other data mining competition that involved multivariate time series data and allowed to achieved very good score (AUC=0.959). This result presented in this paper seems to validate versatility of the proposed approach, as claimed in the original paper.

As suggested earlier, different features seemed to achieve better results comparing to the previous application of the method. Surprisingly, the same basic classifier, namely Random Forest seemed to perform consistently better over other classifiers – the same result was observed in the previous competition.

It is believed that the result presented here provides empirical evidence that the developed approach can be easily generalized to similar problems for which multiple measurements in form of time series are available.

Bibliography

1. Breiman L.: *Random Forests*. "Machine Learning", 2001, No. 1 (45), pp. 5-32.
2. Boullé M.: *Tagging Fireworkers Activities from Body Sensors under Distribution Drift*, Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, 2015, pp. 389-396.
3. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H.: *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, 2009, No. (1)11.
4. Hall M.A.: *Correlation-based Feature Subset Selection for Machine Learning*, Hamilton, New Zealand 1998.
5. Hanley J.A., McNeil B.J.: *A method of comparing the areas under receiver operating characteristic curves derived from the same cases*. "Radiology", 1983, No. (3)148, pp. 839-843.
6. Janusz A., Krasuski A., Stawicki S., Rosiak M., Slezak D., Nguyen H.S.: *Key risk factors for Polish State Fire Service:* A Data Mining Competition at Knowledge Pit, Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, 2014, pp. 345-354.
7. Meina M., Janusz A., Rykaczewski K., Slezak D., Celmer B., Krasuski A.: *Tagging Firefighter Activities at the Emergency Scene: Summary of AAIA-15 Data Mining Competition at Knowledge Pit*, Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, 2015, pp. 367-373.
8. Sikora M., Sikora B.: *Improving prediction models applied in systems monitoring natural hazards and machinery*. "International Journal of Applied Mathematics and Computer Science", 2012, No. 2(22), pp. 477-491.
9. Sikora M., Sikora B.: *Rough natural hazards monitoring*, Rough Sets: Selected Methods and Applications in Management and Engineering 2012, pp. 163-179.
10. Zagorecki A.: *A Versatile Approach to Classification of Multivariate Time Series Data*, Proceedings of the 2015 Federated Conference on Computer Science and Information Systems 2015, pp. 407-410.

The paper was reviewed by two independent reviewers.