# Prediction of Methane Outbreaks in Coal Mines from Multivariate Time Series Using Random Forest

Adam Zagorecki[1,2](✉)

[1] Defence Academy of the United Kingdom, Shrivenham SN6 8LA, UK
[2] Centre for Simulation and Analytics, Cranfield University, Bedford, UK
a.zagorecki@cranfield.ac.uk

**Abstract.** In recent years we have experienced unprecedented increase of use of sensors in many industrial applications. Examples of such are Health and Usage Monitoring Systems (HUMS) for vehicles, so-called intelligent buildings, or instrumentation on machinery in order to monitor performance, detect faults and gain insights in operational aspects. 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. Unfortunately, that collected data requires further analysis in order to provide useful information to the decision makers who want to reduce costs, improve safety, etc. Such analysis proved to be a challenge, as there are no generic methodologies that allow for automating data analysis and in practice costs required to analyze data are prohibitively high for many practical applications. This paper is a step in a direction of developing generic methods for sensor data analysis – it 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 presented application relates to prediction of methane concentrations in coal mines based on time series data from various sensors. The method was tested within the framework of IJCRS'15 data mining competition and resulted in the winning model outperforming other solutions.

## 1 Introduction

In this paper I present an application of a generic approach to classification of multivariate time series data proposed in [1]. This approach was developed and evaluated in the context of the 2015 AAIA Data Mining Competition, where it led to the second highest score. In this paper I present the application of this approach to another data mining competition involving multivariate data series: the *IJCRS'15 Data Challenge: Mining Data from Coal Mines*. The presented solution resulted with the winning entry.

During the recent decade affordable and reliable sensors capable of collecting large amounts of data has become popular in many applications including

Y. Yao et al. (Eds.): RSFDGrC 2015, LNAI 9437, pp. 494–500, 2015.
DOI: 10.1007/978-3-319-25783-9_44

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 consist of sequences of measurements taken over time. With affordable sensors capable of transmitting data over network, multivariate time series data sets are becoming common in many domains. Examples can include vehicle or machinery monitoring, sensors from smartphones or sensor suites installed on human body. Because of the nature of time series, the collected measurements typically not directly exploitable – as the measurements consists of typically a large number of data points, the data is noisy, and requires further analysis in order to identify or discover interesting patterns that can be exploited by users. It is well recognized that the process of 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 in coal mines.

In the application presented in this paper the data consisted of measurements taken by various sensors installed in a coal mine and machinery operating in that coal mine. They 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 min. This knowledge would potentially enable extra warning time before methane warning level is exceeded and can be used to take preventive actions.

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

## 2 The Competition Task

This paper describes a solution to the IJCRS'15 data mining competition which was organized using the Knowledge Pit competition platform [3]. The objective of the competition was to gain insight into dependencies between cutter loader (mining machinery) performance 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 the short future

(3 to 6 min) based on sensors readings from multiple sensors. For this purpose a commercial off-the-shelf body sensor suite was used to generate the data.

### 2.1 Data

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

Each record consisted of 28 time series. Each time series was composed of exactly 600 data points and corresponded to a 10 min 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. The task was to predict methane level exceedance in the future at three sensors. The target variables were three binary variables (threshold exceeded or not) for three selected methane sensors. The true state of a target variable indicated that a methane warning level was exceeded within 3 to 6 min after the end of corresponding the time series.

The 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.

### 2.2 Evaluation

The evaluation of the results was performed using the Area Under the ROC Curve (AUC) measure concept. It was possible, because the target variables were defined in for 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 variables as $AUC_i$. The final score was an average of three individual scores:

$$AUC_{total} = \frac{1}{3}\sum_{i=1}^{3} AUC_i.$$

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.

## 3 Solution Overview

In this section I present an overview of the solution to the competition task. The method I used was based on the method developed for the AAIA’15 data mining competition that is described in [1]. The only difference is that for IJCRS’15 competition I did not use feature selection step. I decided not to use feature selection step as for the this competition feature selection resulted in inferior

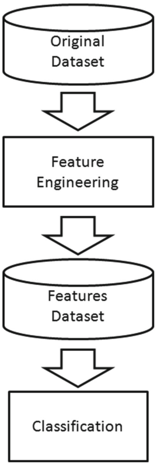

**Fig. 1.** The outline of the basic steps used during the competition.

results and proved unnecessary. The basic steps in the used method are presented in Fig. 1.

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 Sect. 4.

For each of the three target variables I decided to create a separate classifier that made a binary decision. In this way I had to learn three separate classifiers. 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 I used Random Forest, which allowed to compute the probability of class assignment.

### 3.1 Derived Time Series

I decided to expand the original set of time series used 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 $y(t)$, I produced two derived time series $d_1(t)$ and $d_2(t)$ which were:

- $d_1(t) = x(t) - y(t)$ – simple difference between corresponding measurements
- $d_2(t) = \frac{x(t)-y(t)}{x(t)}$ – relative difference between corresponding measurements

The number of derived time series created was 52.

## 4 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 can be derived from individual time series are simple statistics (e.g. mean, standard deviation), more complex features can be derived from more than one time series (e.g. correlation coefficient between two time series). In the course of competition I did a lot of experimentation with different features. I was using feature selection algorithms implemented in Weka software [4] to identify most informative features. As the result of this analysis I put special emphasis on features related to maximal or minimal values, as those seemed to be most informative, at least according to feature selection algorithms. I would like to note, that I used the feature selection to inform feature engineering only – I did not use feature selection to actually select features for classification – I used all generated features for classification task.

### 4.1 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 2214 features – 756 from the original time series and 1458 from derived time series.

### 4.2 Correlations

Finally, I decided to add correlation coefficients between time series. Additional parameters were derived from cross-correlations (those included auto-correlations) 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 $y(t)$ 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 I allowed $x(t)$=$y(t)$. The total number of features in the winning set was 4914.

## 5 Classification

I used Random Forest [6] implemented in Weka software [4] as the basic classifier. I did experimented with other classifiers such as Neural Networks, Logistic Regression, Support Vector Machines, and others, 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. I experimented with different numbers of features per tree and for the particular feature set the numbers between 60 and 100 features seemed to work well. For the best score I could achieve, each of three Random Forest classifiers had 1000 trees. The number of features for each tree was limited to 80.

## 6 Conclusions

In this paper I presented winning solution for the IJCRS'15 Data Mining Competition. The approach is a slightly 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. This result 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.

I believe 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.

## References

1. Zagorecki, A.: A Versatile Approach to Classification of Multivariate Time Series Data. In: The Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (2015, to appear)
2. Meina, M., Janusz, A., Rykaczewski, K., Ślęzak, D., Celmer, B., Krasuski, A.: Tagging firefighter activities at the emergency scene: summary of AAIAâĂŹ15 data mining competition at knowledge pit. In: Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (2015)
3. 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. Federated Conference on Computer Science and Information Systems (FedCSIS) 2014, pp. 345–354 (2014) doi:10.15439/2014F507
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations, vol. 11(1), pp. 10–18 (2009)
5. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand (1998)
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)