Knowledge discovery and anomaly identification for low correlation industry data

Zhe Li[1] , Jingyue Li[1]

Abstract. With the development of information technology, industry data is increasingly generated during the manufacturing process. Companies often want to utilize the data they collected for more than the initial purposes. In this paper, we report a case study with an industrial equipment manufacturer to analyze the operation data and the failure records of the equipment. We first tried to map the working condition of the equipment according to the daily recorded sensor data. However, we found the collected sensor data is not strongly correlated with the failure data to capture the phenomenon of the recorded failure categories. Thus, we proposed a data driven-based method for anomaly identification of such low correlation data. Our idea is to apply a deep neural network to learn the behavior of collected records to calculate the severity degree of each record. The severity degree of each record indicates the difference of performance between each record and all other records. Based on the value of severity degree, we identified a few anomalous records, which have very different sensor data with other records. By analyzing the sensor data of the anomalous records, we observed some unique combinations of sensor values that can potentially be used as indicators for failure prediction. From the observations, we derived hypotheses for future validation.

Keywords: Knowledge discovery, Anomaly identification, Data correlation

1 Introduction

Recent advances in information and communication technologies have accelerated the application of automated and systematic monitoring systems in the manufacturing industry [1]. In the past few years, the dimensionality of monitoring data sets in the manufacturing industries has severely increased [2,3]. Hence, issues about how to leverage those monitoring data to enhance the reliability and availability of the equipment are getting more and more significant [4].

In this work, we want to help an industrial equipment manufacturer to analyze the data they collected, namely the operation data and the failure log of the equipment. Our initial target is to predict potential failures and working conditions of the target equipment from its daily collected monitoring data, which has a total number of 197 parameters. We first applied a classification model to identify the difference between failure

[1] Zhe Li (✉) and Jingyue Li
Department of Computer Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway
{zhel, jingyue.li}@ntnu.no

and normal records. According to our analysis results, the applied data-driven model could not separate the failure records from the nonfailure records. We assume that the sensor data have no strong correlation to the failures. We generated a new parameter to indicate failure conditions and further validated and confirmed our assumption that available sensor data is not strong enough to capture the phenomenon of failures. Actually, how to extract valuable information and discover useful knowledge from ***low correlation*** data is a common dilemma in many practical applications [5]. To fill the gap, we proposed a method to discover knowledge and identify anomaly for low correlation data. The applied data-driven model is constructed through a fully connected deep neural network since this structure has an excellent performance in discovering information and knowledge about failures [6]. The idea is to make the constructed neural network learn the behavior of collected samples and output the severity degree of each sample. The severity degree can indicate the difference of performance between individual record and all other records. From the severity degree, we can identify anomalous records. Through analyzing the sensor data of the identified anomalous records, we could acquire knowledge about which sensor data could be possible indicators or predictors of failures. Although the study is based only on one equipment data, knowledge acquired from the case study could help us derive hypotheses for validation by using other similar equipment.

The rest of this paper is organized as follows: Section 2 explains the process of data correlation analysis and validation. Section 3 details the applied method for knowledge discovery and anomaly identification for low correlation data. Discussion and conclusions are summarized in the last section of this paper.

2 Data correlation analysis

The data we used during the research is collected from the equipment of an industrial equipment manufacturer[2]. The primary datasets leveraged during analysis includes two parts: fault records and sensor data from the monitoring system. However, the sensor data is mainly collected to help the user understand the working condition of the equipment, instead of to indicate fault information. The target of our study is to discover the potential correlation between the two databases and knowledge about impending failures.

2.1 Data integration and normalization

According to the measurement system, the sensor database includes 247269 records within 2931 timestamps. The number of recorded parameters varies with years, from 57 parameters in 2009 to 197 parameters in 2017. Forty-two monitoring parameters, which have been collected all the time during the sampling period, are selected as the observation units for further analysis. The leveraged monitoring data mainly include a total number of starts, running time, load, voltage, and so on.

[2] Due to Non-Discloser Agreement, we are not allowed to give detailed information of the equipment and the company in the paper.

There are 538 records from 2009 to 2017 in the provided fault dataset. The collected data includes 315 days and 367 timestamps. Several failures may happen in one timestamp, and several timestamps may be recorded in one day. During the sampling period, most of the recorded failures are about faults in multi-hoisting.

To integrate monitoring data and failure information, we used the recorded timestamps in each database as connections. Since both monitoring data and failure information are necessary for fault identification, we included only the records which have been recorded in both datasets as observations in the analysis.

The number of valid records is 2931 after the merge, in which 307 records are labeled as failures, and 2624 ones are not labeled as failures. As mentioned above, the number of collected timestamps in monitoring dataset and failure dataset are 2931 and 367, respectively. However, there are 60 records in fault dataset which are not recorded in monitoring dataset. Thus, we discarded these 60 records without monitoring information in the analysis.

To improve the performance of data mining and avoid potential inconvenience, we applied standard normalization to adjust values measured on different scales to a notionally common scale. The parameter P_i after normalization is P_i' , which is shown in Equation (1):

$$P_i' = \frac{P_i - P_i^{mean}}{P_i^{std}} \tag{1}$$

Here, P_i^{mean} and P_i^{std} are the mean value and standard deviation of the parameter P_i, respectively.

During normalization, we found that the standard deviations of seven parameters are zero or very close to zero, which means those parameters rarely changed during the sampling period. We removed these parameters from the merged dataset since constant values hold no meaning for condition monitoring.

As mentioned above, since the currently available monitoring data is used for operation management, there is probably no direct connection between the available monitoring data and failure information. Therefore, our first research step focused on answering whether the collected monitoring data is sensitive or strong enough to predicate impending failures.

2.2 Data analytics for impending failure prediction

In this section, we will introduce the process and test result of impending failure prediction. Our target of this step of analysis is to leverage the collected monitoring data and data-driven models to map the recorded fault conditions. If the collected monitoring parameters are sensitive or relevant enough to identify the recorded failures, we can use them to predict whether there would be an impending failure. Thus, the problem is transformed into a classification issue with two groups, i.e., a non-fault group and a fault group. As mentioned above, our dataset includes 307 records under impending failure condition and 2624 records which are not labeled as failures. To balance the number of samples in both classes, we expanded the number of failure samples by repeatedly used them during the training stage.

The applied data-driven model is established through the fully connected deep neural network with seven layers, Leaky Relu is used as activation functions of hidden layers, and SoftMax is used as activation functions of the output layer. Since the dimension of inputs is 35 (42 selected parameters minus seven constant values), the number of nodes in hidden layers of the constructed deep neural network is 64, 32, 32, 16, 16, 8, 2 (i.e., 64 nodes in the first layer, 32 nodes in the second and third layer, 16 nodes in the fourth and fifth layer, 8 nodes in the sixth layer, and 2 nodes in the last layer) to learn and represent the inputs data smoothly. We selected Adam as the optimizer and categorical cross-entropy as the loss function due to their broad applicability. The maximum number of training epochs and dropout rate are 2000 and 0.3, respectively, to avoid overfitting. Batch size has been set as 32 to accelerate the training process. Fig. 1 shows the training and validation error with training epochs. Fig. 2 illustrates the prediction result, in which values above 0.5 in y-axis can be considered as identified failure.

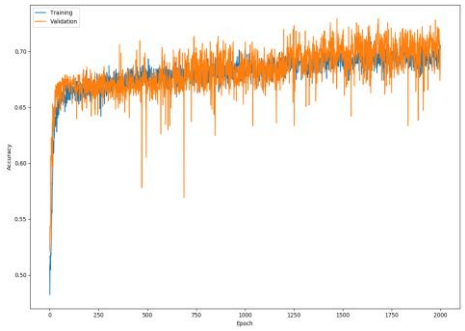

Fig. 1. Training errors with epochs

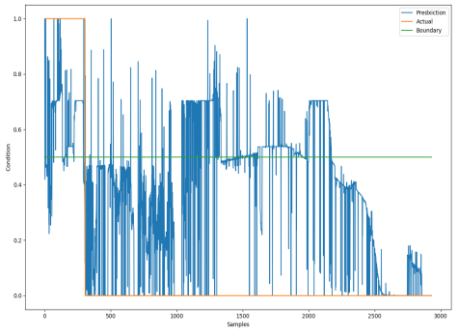

Fig. 2. Impending failure prediction

Since we used SoftMax as the final output layer, records with higher values in y-axis are more likely to indicate a failure condition. According to the result, we can notice that only some extreme normal and failure conditions can be identified. Most records have the prediction result close to 0.5, which means those records are difficult to be separated by the sensor data. We assume the main reason is that current sensor data is not sensitive enough to capture the phenomena of failures, or not strong enough to identify the impending failures.

2.3 Correlation validation

To validate this assumption, we set up a new experiment with more parameters, generated from failure information with certain random fluctuations (0.7 – 1 for failure records and 0-0.3 for nonfailure records). The generated parameters can be considered as direct indicators, which are sensitive to recorded failures and can capture the phenomena right before an impending failure. Fig. 3 shows the result of failure prediction with the generated parameters using the same deep neural network. According to the result in Fig. 3, the collected parameters are sensitive enough to the failures, and the data-driven model can identify the failure conditions. Results of Fig. 3 confirms that the original sensor data cannot capture the phenomenon of impending failures directly.

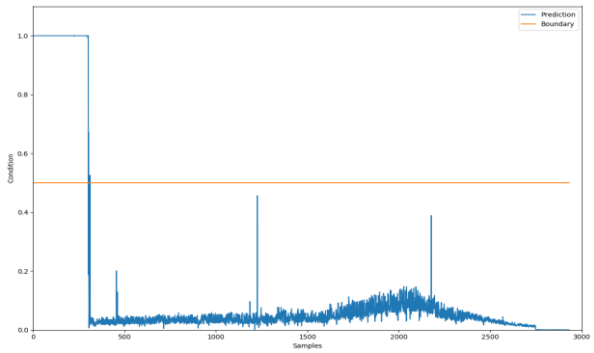

Fig. 3. Failure prediction with generated parameters

3 Knowledge discovery and anomaly identification

As the available sensor data is not sensitive enough to capture the phenomena of failures, we tried to leverage the data to discover possible useful knowledge hidden in the data. Since most of the records are collected without failures, i.e., only 307 in 2931 records are labeled as a failure, our core idea is anomaly identification. The idea is to make a data-driven model learn behaviors of the equipment first. The second step is to give scores to each record to describe the degree of difference from other records. The high-level analysis process is shown in Fig. 4.

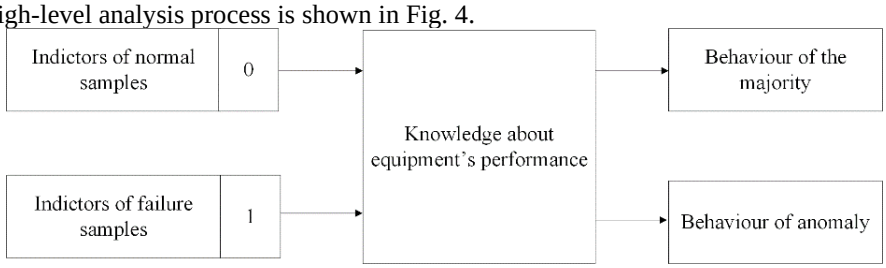

Fig. 4. Knowledge discovery from low correlation data

During data calibration, we labeled all the failure records as "1" and the nonfailure ones as "0". Therefore, the records with higher scores are more likely to have impending failures or anomaly condition since their behaviors are different from others. The target is to identify records with abnormal behaviors [7], which are the records having very different sensor data with other records. The anomalous records may or may not have been labeled as failure ones in the original dataset.

The applied data-driven model for anomaly identification is very similar to the failure identification model, which is deep neural works with seven fully connected layers. The difference from the failure identification model is that the final layer is replaced with a regression model to evaluate the anomaly degree and output severity degree. Fig. 5 shows the evaluation results of the trained network with a hyperbolic tangent (Tanh) as the activation function of the final output layer.

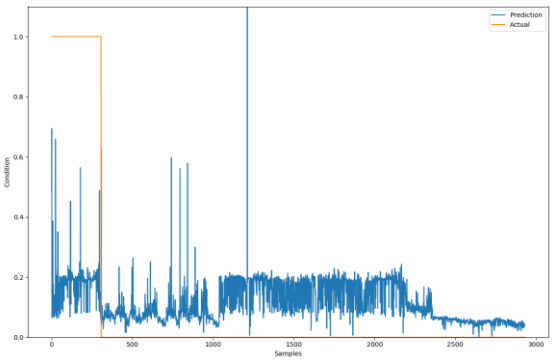

Fig. 5. Evaluation results of Tanh function

During the evaluation phase, the first 307 records are the ones with failures, and the rest records are the nonfailure ones. In Figure 5, the red line indicates the actual severity degree, and the blue line represents the prediction result. According to the evaluation result, the severity degree of most records is around 0.2. Thus, we selected 0.22 as the threshold. Twenty-nine out of 2931 records are identified as the anomalous ones from the analysis. Among the anomalous 29 records, 15 records are labeled as nonfailure ones in the original dataset, and 14 are labeled as failure ones. According to the results, all records can be divided into four categories:

- **Category 1**. 2609 nonfailure records are also identified by our data model as normal. Those samples represent normal behaviors without failures.
- **Category 2**. 293 records, which are recorded as failure ones in the original dataset, are not identified as anomalous by our data-driven model. The data-driven model cannot identify any difference between these records labeled as failure and records labeled as nonfailure ones. We analyzed the sensor data of many of those records in details and compared their sensor data with the ones labeled as nonfailure and find that their sensor data are very similar to the sensor data of the nonfailure ones. ***As these records cover 95% of all the records labeled as failures, their sensor data may hide some interesting correlations between the sensor data and the failures. Such interesting correlations probably exist in some data but are not statistically significant.*** This probably explains why our initial analysis reported in Section 2 did not find strong correlations between the sensor data and the failure.
- **Category 3**. 14 records, which are recorded as failures in the original dataset, are identified by our data-driven model as anomalies. Those samples are recorded with failure and have very different sensor data with other records. Thus, they are captured by our data-driven model. The 14 records in Category 4 are the real interesting records to be analyzed further, because such record may contain implicit knowledge that can potentially explain the reasons for the failures and can also possibly help us identify real indicators from the sensor data to predicate failures. Thus, we analyzed the sensor data of these records in details. Four out of the 14 records have similar conditions that one or several vital

parameters such as “total running time,” “Remaining Safe Working Period of the brake in percentage,” and “a total number of starts” are recorded with extremely high or low values. Those unusual high or low values indicate sensor failures of the equipment. Six of the records are identified as an anomaly because the “*actual loads*” of the equipment are higher than the average value, but values of other parameters are different with most samples with high actual loads. Thus, our data-driven model identifies them as anomalies since their conditions are different from most. Among these 6 records, 4 records have high values in “*actual load*,” while one or several “*line voltages*” are lower or close to the average values. As these 4 records are also labeled as failures in the original dataset, these records may make one out of many possible reasons for the failures stand out. Thus, we can hypothesize that “***the target equipment under high load without enough line voltages is more likely to have an impending failure***.” For the rest 4 records, our visual inspection could not find obvious abnormal of the value of their individual sensor parameter. As each record has 35 sensor parameters, it is possible that some complex combinations of sensor parameter values make them very different from the other records. More domain knowledge and more data are needed to understand these 4 records in depth.

- **Category 4**. Fifteen records, which are not categorized as failed one in the original data set, are identified as abnormal by our data model. By analyzing the sensor data of the 15 records in depth, we found 4 out of the 15 records have sensor failures that were not recorded or noticed by the users. These indicate sensor errors. However, due to unknown reasons, the sensor errors do not lead to actual failures or the actual failures are overlooked. There are 9 records which have high “*actual loads*” as the 4 records, which also have high “*actual loads*,” in Category 3, Although these 9 records are not labeled as failures in the original dataset, their sensor data are very similar to the 4 failure records with high “*actual loads*.” That is probably why these 9 records are also identified as abnormal by our data-driven model. Again, there are 2 records we cannot figure out how different their individual sensor parameter values are from other records. We need more in-depth domain knowledge and more data to explain the reasons why our data-driven model classifies these 4 four records as abnormal.

4 Discussion and conclusion

According to the result of failure identification in Section 2, the currently available sensor data are not strong enough to predicate failures. Thus, we change our research focuses on anomaly identification and proposes a method to evaluate severity degree by comparing the behavior of each record with the records which are recorded as non-failures. As shown in Section 3, we first established a data-driven model to evaluate the severity degree of each record. The core idea is to train the model and make it learn the behaviors of the majority. Thus, the evaluated severity can indicate the degree of anomaly condition of each record compared with all other records. A record with high severity is more likely to have anomaly behaviors.

The proposed anomaly identification method can identify anomaly behaviors of the target equipment and obtain hypotheses about machine fault from low correlation data environment. In this case study, our approach filtered out most of the records which are labeled as failures in the original dataset but are not able to differentiate themselves from the nonfailure records by inspecting the sensor data. Our approach managed to find out 4 records which have extremely high or low sensor values and are labeled as failures. These 4 records indicate that sensor error is probably one reason for failure. Our approach also highlighted other 4 records which have high "*actual load*" but "*line voltages*" and are labeled as failures. Such 4 records may indicate that the parameters "*actual load*" and "*line voltages*" can possibly be used as indicators for predicting some categories of failures.

The limitation of the study is that the proposed method is only applied and tested with data from one equipment. We, therefore, need to validate the method proposed in this study and the hypotheses identified from this study with data from several similar types of equipment.

Acknowledgment

The work described in this article has been conducted as part of the research project CIRCit (Circular Economy Integration in the Nordic Industry for Enhanced Sustainability and Competitiveness), which is part of the Nordic Green Growth Research and Innovation Programme (grant numbers: 83144), and funded by NordForsk, Nordic Energy Research, and Nordic Innovation.

References

1. Lee J., Bagheri B., Kao H.-A. (2015) A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters 3:18-23
2. Lee C. K. M., Zhang S,Z,, Ng K. K. H. (2017) Development of an industrial Internet of things suite for smart factory towards re-industrialization. Advances in Manufacturing 5 (4):335-343
3. Bodrow W. (2017) Impact of Industry 4.0 in service oriented firm. Advances in Manufacturing 5 (4):394-400
4. Li Z., Wang Y., Wang K.-S. (2017) Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario. Advances in Manufacturing 5 (4):377-387. doi:10.1007/s40436-017-0203-8
5. Khan A., Turowski K. (2016) A Survey of Current Challenges in Manufacturing Industry and Preparation for Industry 4.0. Paper presented at the Intelligent Information Technologies for Industry, Cham,
6. Li Z., Wang Y., Wang K. (2019) A deep learning driven method for fault classification and degradation assessment in mechanical equipment. Computers in Industry 104:1-10
7. Chandola V., Banerjee A., Kumar V. (2009) Anomaly detection: A survey. ACM Computing Surveys 41 (3):15