Advanced Correlation-Based Anomaly Detection Method for Predictive Maintenance

Pushe Zhao, Masaru Kurihara, Junichi Tanaka
Center for Technology Innovation - Electronics
Hitachi, Ltd., Research & Development Group
Tokyo, Japan

Tojiro Noda, Shigeyoshi Chikuma, Tadashi Suzuki
Internet of Things System Development Dept.
Information & Control Systems Div.
Hitachi Power Solutions Co., Ltd.
Ibaraki-ken, Japan

***Abstract*—Variations in sensor data collected from equipment have been widely analyzed by using anomaly detection methods for predictive maintenance. Our experience shows that correlations between sensors effectively predict failures because the correlations usually reflect the status of equipment with higher sensitivity. In this paper, we present a method that exploits correlations between sensors for pre-processing and enables anomalies to be detected using both sensor data and correlations. The method was evaluated by applying it to compact electric generators, and the results showed it detected anomalies more accurately than when only sensor data were used. This method is expected to predict failures earlier and reduce the cost of downtime and maintenance.**

Keywords-predictive maintenance; multivairate time series; correlation coefficient; anomaly detection; electric generator

I. INTRODUCTION

Predictive maintenance (PdM) has become widely used because it can improve productivity and save operating costs in various industries, such as manufacturing, infrastructure, healthcare, and energy. PdM involves using data analysis to maintain electronic devices, manufacturing machines, and infrastructure systems more effectively and efficiently than conventional ways, such as corrective maintenance (CM) and preventive maintenance (PM) [1, 2].

In CM, machines (in this case, electric generators) in factories are repaired after a failure occurs, and the users have to purchase electricity at a higher cost during the downtime of their electric generators. On the other hand, in PM, engineers go to check these electric generators and replace components in accordance with a fixed schedule (e.g., once a month). Although PM can avoid most of the breakdowns of electric generators in CM, PM usually leads to higher cost due to more frequent maintenance. Furthermore, the time taken for scheduled maintenance actually leads to reduced production time and productivity. To avoid these disadvantages, PdM predicts potential failures by detecting the degradation of performance of electric generators, and maintenance can be planned prior to the failure on the basis of the prediction results [3-5]. During this process, schedules of engineers can be arranged appropriately, and necessary components for replacement can be prepared in advance. In this manner, high production efficiency can be achieved by optimizing the maintenance schedule.

In this work, we focused on developing a PdM solution for compact electric generators, and its features are summarized in Table I. The target is to predict potential failures before they happen, which makes it possible to analyze the cause of a potential failure and make a maintenance plan. Although a model of a target generator and knowledge about the generator can also be used for predicting failures, only sensor data collected from generators were used in this work. There are several ways to predict a failure, such as using a degradation curve and simulation results. An anomaly detection method was adopted in which anomalous sensor data are detected as a sign of potential failure by comparing them with normal sensor data. The output of this process is an anomaly alert, and the time between the alert and failure (which will occur if no maintenance is conducted), which is defined as the lead time of prediction, is used to evaluate the effectiveness of a PdM solution. The PdM process for equipment in this work is illustrated in Fig. 1.

TABLE I. SUMMARY OF PDM IN THIS WORK

Target	Early prediction of potential failure
Input	Existing sensor data
Method	Anomaly detection based on statistical analysis
Output	Alert of anomaly

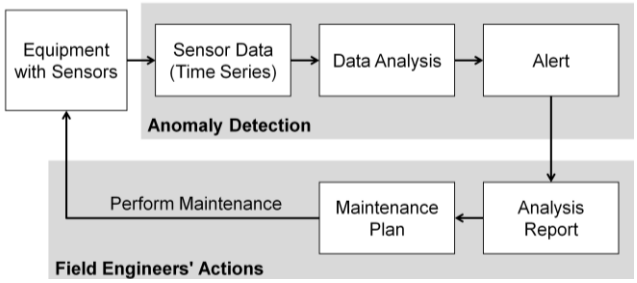

Figure 1. Process of predictive maintenance in this work

978-1-5090-0382-2/16/$31.00 ©2017 IEEE

978-1-5090-5710-8/17/$31.00 ©2017 IEEE

There are basically two ways to analyze the data for anomaly detection: statistical and machine learning methods [6, 7]. Among them, we specifically consider methods based on correlation analysis of the sensor data to be effective for analyzing data of complex equipment for three reasons. First, in a machine, the sensors are usually correlated, and these correlations physically reflect the mechanisms and operating conditions of the machine. Second, unlike sensor data, a correlation usually changes abruptly when an anomaly appears. This makes the correlations more sensitive to the status variation of equipment. Third, the correlation analysis is comparatively simple to implement, which makes it possible to be used as a high-speed (real-time) but low-cost solution.

A correlation-based anomaly detection method has been developed by Zhong et al. [8]. Correlation coefficients (CC) between every pair of sensors are calculated and transformed into a latent correlation vector. Then a probabilistic model is applied to the vector to detect anomalies. The method has been evaluated by using a real flight dataset and showed better detection accuracy. However, there are still two challenges. First, all the CCs between sensors are used for analysis, which may cause a dimensionality problem when many sensors exist. Second, the probabilistic model needs to be optimized to detect anomalies more accurately. Therefore, in this paper, we propose an improved correlation-based method that overcomes these challenges and try to apply it to a compact electric generator by designing processing functions necessary to use the method.

To avoid the dimensionality problem, we designed a function to select CCs that are estimated to be effective at anomaly detection. The idea is that not all the sensors correlate, and only the steady ones are suitable for anomaly detection. To successfully use the method on a compact electric generator, many other considerations are necessary. For example, in this work, sensor data are preprocessed to avoid the effect of outliers, noise, multiple operating conditions, and trends in sensor data.

The rest of the paper is organized as follows. Section II explains the improved method of anomaly detection in detail. Section III details how the proposed method was applied to a compact electric generator gives the experimental results for evaluation. Section IV concludes this paper and discusses further analysis and future work.

II. METHOD

A. Correlation Coefficient for Predicting Failures

The correlations between sensors in a complex machine usually reflect the state of the machine. Although a correlation may either be strong or weak, an actual correlation in a machine operating in a steady-state is expected to obey a statistical distribution, such as a normal distribution. Therefore, when a change occurs in the machine, a correlation may deviate from its original distribution. The correlation can be considered as complementary to sensor data. Although various methods [9] measure the correlation between two variables, the Pearson correlation coefficient is adopted in this paper. Its definition is shown as (1), where x, y are data of two sensors, and r is the Pearson correlation coefficient value, ranging from -1 to 1.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

Suppose a machine has two sensors that correlate when the machine is operating in a normal state. By monitoring variation of sensor data, some failures can be predicted by detecting anomalies. For example, by setting a threshold to data of sensors A and B, an anomaly can be detected, as shown in Fig. 2. Since the sensor B data increase gradually, the value of B exceeds the threshold after a certain time. When the correlation between A and B is monitored, anomalies can be detected earlier because when the sensor B data start to increase, the correlation between A and B will change correspondingly. Because the change is abrupt, it can be detected more easily. In fact, the correlation can be considered as a logistic variable in this case, which makes it more sensitive than sensor data. The correlation-based method can detect anomalies better than the method using only sensor data. However, sensor data may be more sensitive to certain kinds of failures. Therefore, we actually combined CCs with sensor data to achieve a better performance.

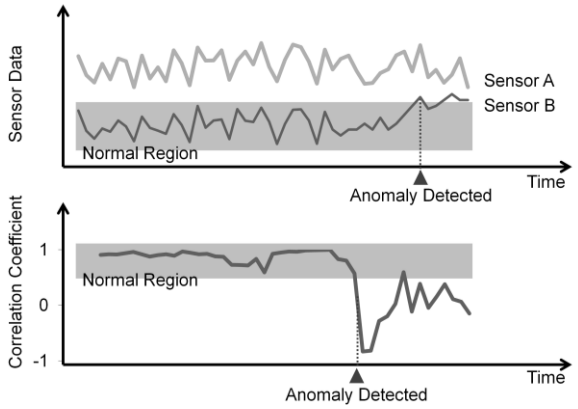

Figure 2. Illustration of anomaly detection using correlation coefficient

B. Calculation of Correlation Coefficient

Sensor data are time-series data, and the CCs are calculated using data of a certain length. Figure 3 shows an example of the calculation. A time series of a CC between sensors 1 and 2 is calculated. For real sensor data, this calculation requires more data than shown in the example and needs to follow the statistical requirements for calculating CC.

According to the characteristics of sensor data, other methods for assessing correlations can also be used. For example, advanced methods, such as MIC [9] and dCor [11], can be used to achieve more accurate correlation assessment, which can probably improve the performance of anomaly detection on complex equipment.

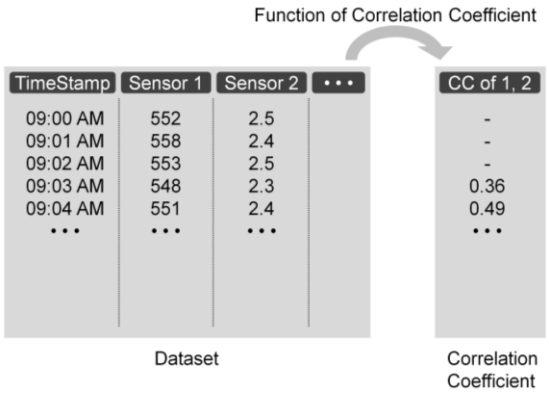

Figure 3. Calculation of correlation coefficient (time series)

C. *Selection of Correlation Coefficient*

It can be observed that the number of CCs increases quickly with the number of sensors. When the number of sensors is large, the CCs may become difficult to process. For example, many anomaly detection algorithms suffer from the high-dimensionality problem, which leads to large computational effort and low accuracy. In the work of Zhong et al. [8], all the CCs are calculated and arranged to form a vector called a latent correlation vector without further processing. However, not all the CCs are useful in anomaly detection. The CC between two sensors that are not correlated probably gives little information and is just noisy values around zero. On the basis of this assumption, we designed a correlation selecting function to select effective CCs in accordance with their strength and stability.

D. *Vector Quantization Clustering*

To detect anomalies in CC data, a function is required that creates a model of normal data. In this work, we adopted a clustering method to process sensor data and CC data at the same time. There are two reasons for this. First, as already explained, both types of data are complementary and combining them can lead to higher performance. Second, the CC data are also time series data, which are very convenient to use as additional sensor data without making big changes to the anomaly detection part. The CC data are aligned with sensor data in accordance with timestamps. Both data are merged into a new dataset as shown in Fig. 4.

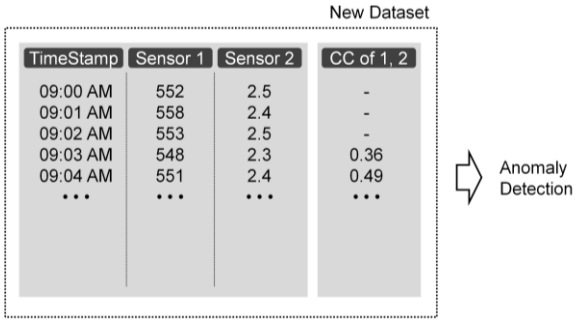

Figure 4. Merging of correlation coefficient with sensor data

Zhong et al. [8] analyzed the CCs by using a Gaussian-distribution model, which requires the assumption of Gaussian distribution. In this paper, we adopted a more general method called vector quantization clustering [10], which generates a model of normal data using many clusters. It can analyze data requiring no assumptions about the distribution of the data. To tell whether new data are anomalous or not, the distance (called the anomaly measure) between the data and those clusters is calculated. When the distance exceeds a certain threshold, the data are considered as anomalous, as shown in Fig. 5.

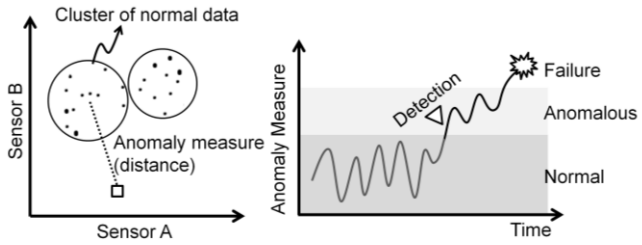

Figure 5. Anomaly detection using vector quantization clustering

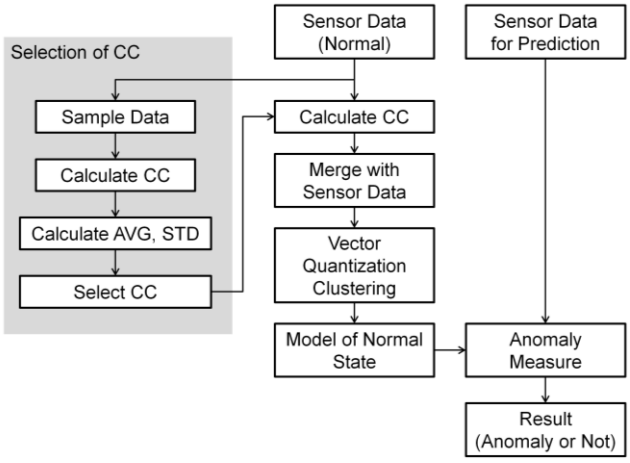

Figure 6. Process of proposed anomaly detection method. CC, AVG, and STD stand for correlation coefficient, average value, and standard deviation value.

The process of the proposed method is summarized in Fig. 6. First, part of the data is extracted from the dataset to select CCs. Then the time series of selected CCs are calculated and added to the original dataset of sensor data. The new dataset is processed by a clustering algorithm to generate a normal model. Lastly, data for prediction are applied to the model to detect anomalies. To apply it to actual equipment, processing functions need to be designed, which are explained below.

III. EXPERIMENT

A. *PdM of compact electric generator*

This section describes using the proposed method for PdM of compact electric generators to evaluate its effectiveness. For evaluation, we use the index of lead time, which is defined as the time period between the prediction (alert) and the

occurrence of failure. A longer lead time is preferred because there will be sufficient time for engineers to find the cause and prepare a maintenance plan. Furthermore, if an anomaly is detected at an early stage, the anomaly is unlikely to have done much damage to the machine, which probably means lower maintenance cost.

Twenty sensors are installed on the generator used in this experiment, such as temperature sensors and pressure sensors, which send data to a computer at a certain rate. Generators operate differently, for example, 24 hours per day as high-load, or 10 hours per day as varying load. Therefore, the method needs to be able to deal with generators operating under different conditions. Most sensor data are normal data, while only a small amount of anomalous data of failures is recorded. Therefore, we can generate the anomaly detection model using only the normal data in this case. In the following, the application is explained in the order of preprocessing, parameter settings, and anomaly detection.

B. Data Processing

Data must be preprocessed as preparation for the proposed method to perform as expected. The time series may have data of different modes, for example, the output power of a generator may be changed in accordance with a generator's load. Since the proposed method works well with steady-state data, a data separating function is designed to extract data of a different mode first and then detect anomalies in the data separately. The data separating function uses thresholds set in advance to determine the mode of data; for example, if expected output power is P, than data falling around P are extracted. Other preprocessing was carried out, such as removing outliers and standardization.

C. Parameter Settings

Parameter settings can greatly affect performance of the method, so this subsection describes the settings of a threshold on an anomaly measure. The settings of a threshold for anomaly detection are shown in Fig. 7. After the anomaly measure (distance) of sensor data is calculated, the score has to be compared with a threshold to determine whether the data are anomalous or not. This threshold is determined by applying the model of normal data to other normal data to find the variation of an anomaly measure. For example, the normal data are divided into two parts at the beginning. Then a model is created by using the first part and applied to the second part. The distribution of the results (anomaly measures) shows the variation of anomaly measures of normal sensor data. By using this distribution, the threshold can be set as shown in Fig. 7.

D. Experimental Settings

Experiments were carried out to evaluate the performance of this method and reveal the physical mechanisms behind the results. Experimental settings are summarized in Table II. One-year sensor data of five generators were used, and 25 known failures were recorded. The proposed method is applied to the data as if they were time-series data following the preprocessing and settings described above. For data of a certain time point, the data two weeks before were used to create the model, and the data are determined to be anomalous or not. Then the data of the next time point were processed. The anomaly detection process continued until one failure was detected. The lead time for that failure was calculated, and the anomaly detection started again from the end of maintenance for that failure. The index for evaluation was the lead time for each failure. If no anomaly was detected before one failure, the lead time was set to zero.

The purpose of the experiment is to evaluate whether using correlations improves the performance of anomaly detection. Therefore, we compared the lead time of anomaly detection when using only sensor data and using both sensor data and CCs. The clustering function, process, and parameter settings were the same for both cases.

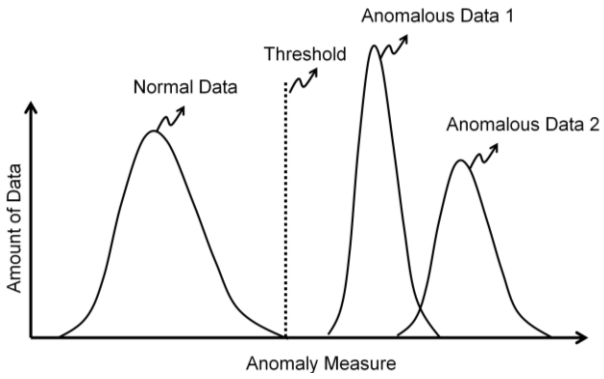

Figure 7. Setting of threshold for anomaly detection

The process of anomaly detection application on a compact electric generator is summarized in Fig. 8.

Figure 8. Process of anomaly detection application

TABLE II. SUMMARY OF EXPERIMENTAL SETTINGS

Machine	Compact electric generator
Number of machines	5
Number of failures	25 (different kinds of failures)
Number of sensors	20
Evaluation	Lead time (time between prediction and failure)

E. Experimental Results and Discussion

The experimental results are shown in Fig. 9. For each electric generator A~E, average lead times of anomaly detection using only sensor data and using both sensor data and CCs were calculated. The results reveal that the lead times were improved (longer lead time preferred) by analysis on CCs, which means CCs have better sensitivity to some failures. It can also be observed that analysis on CC is not effective for all

kinds of failures. However, since the proposed method used the sensor data and CCs, it took advantage of both.

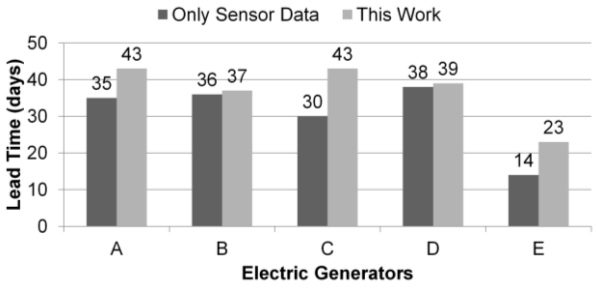

Figure 9. Comparison of lead time w/wo correlation coefficients

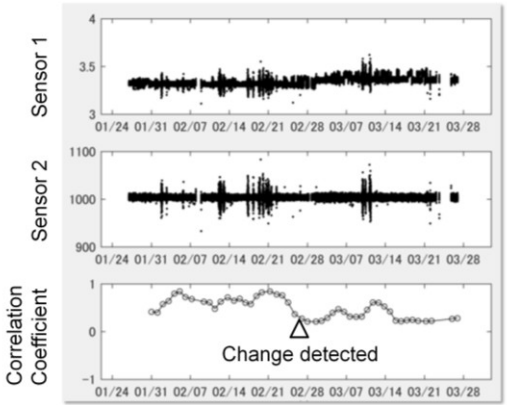

Figure 10. Example of anomaly detection using correlation coefficient

F. Verification

To verify the experimental results, we also tried to figure out the physical mechanisms suggested by the CCs by asking field engineers for help. Since the experimental results show that the CC contributing most to the anomaly detection can be extracted, the first point confirmed by the engineers is whether the CC is related to a failure. The second point is whether the extracted CCs (which may not be related to a failure) are physically meaningful. The results show that about 80% of the selected CCs are considered to probably exist. Although this result is not quantitative, it still shows that the latent correlations that reflect the physical mechanism of a machine can be extracted by using CC analysis.

In fact, the results are more complicated than we expected because real correlations in generators may include more than two sensors. However, we were still able to find simple examples, one of which is shown in Fig. 10. The anomaly detection results reveal that the variation in CC between sensors 1 and 2 shows signs of an anomaly. We extracted the data of sensors 1 and 2 and calculated their CC for observation. Although a change in data of two sensors is not easy to observe, in the time-series of CC, the change is very clear and easy to detect. This shows the effectiveness of CC in predicting failures.

Data variation of CC shown in Fig. 10 also implicates the reason why CC is effective for anomaly detection. When the target generator operated normally the CC varies near a certain value without obvious change. When some anomaly appears, the anomaly can be detected by using sensor data only if the values of sensor data exceed pre-defined threshold. In practice there are many cases that the threshold-based method does not work well. However, in these cases that values of sensor data do not exceed the threshold, anomaly can probably be detected by using CC analysis, because the CCs deviate from normal region clearly and quickly.

From viewpoint of lead time, although the anomaly detection of generator B and D are improved by only one day, the results of generator A, C, and E are improved greatly, as shown in Fig. 9. The improvement strongly depends on the cause of the potential failures. If the anomaly occurs at a part (sub-system) of the generator that are monitored by multiple sensors, the correlations between those sensors will change, and the lead time can be improved. On the other hand, if the anomaly occurs at a part with only one sensor installed, using CC analysis or not may result in almost same lead time.

IV. CONCLUSIONS

An improved correlation-based anomaly detection method has been developed as a predictive maintenance (PdM) solution for compact electric generators. In this method, correlations are selected by using statistical analysis, and anomalies are detected by using both sensor data and correlation coefficients (CCs). The method has been applied on compact electric generators with necessary functions for data analysis. Experimental results showed the proposed method detected anomalies more accurately than conventional methods, which proved its effectiveness. The results were also verified by field engineers to confirm the physical mechanisms suggested by the CCs. This method can be expected to provide better PdM for equipment or factories.

REFERENCES

[1] B. Lu, D. B. Durocher, and P. Stemper. “Predictive maintenance techniques,” *IEEE Industry Applications Magazine* 15.6 (2009): 52-60.

[2] A. Grall, L. Dieulle, C. Berenguer, and M. Roussignol, “Continuous-time predictive-maintenance scheduling for a deteriorating system,” *IEEE Transactions on Reliability*, Vol. 51, No. 2, June 2002.

[3] X. Jin, D. Siegel, B. A. Weiss, E. Gamel, W. Wang, J. Lee, and J. Ni, “The present status and future growth of maintenance in US manufacturing: results from a pilot survey,” *Manufacturing Review* 3, 2016.

[4] D. Goyal, A. Saini, S. S. Dhami, and B. S. Pabla, “Intelligent predictive maintenance of dynamic systems using condition monitoring and signal processing techniques—A review,” In *Advances in Computing, Communication, & Automation (ICACCA)(Spring), International Conference on* (pp. 1-6), April 2016.

[5] E. Zio, "Prognostics and health management of industrial equipment," *Diagnostics and prognostics of engineering systems: methods and techniques*, 2012, pp. 333-356.

[6] H. M. Hashemian and C. B. Wendell, "State-of-the-art predictive maintenance techniques," *IEEE Transactions on Instrumentation and measurement* 60.10 , 2011, pp. 3480-3492.

[7] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: a multiple classifier approach," *IEEE Transactions on Industrial Informatics*, *11*(3), 2015, pp. 812-820.

[8] S. Zhong, H. Luo, L. Lin, and X. Fu, "An improved correlation-based anomaly detection approach for condition monitoring data of industrial equipment," In *Prognostics and Health Management (ICPHM), 2016 IEEE International Conference on* (pp. 1-5). IEEE.

[9] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, *334*(6062), 2011, pp. 1518-1524.

[10] T. Suzuki, T. Noda, H. Shibuya, and I. H. Suzuki, "An Anomaly Detection System for Advanced Maintenance Services," *Hitachi Review*, *63*(4), 178, 2014.

[11] G. J. Székely, and M. L. Rizzo, "Brownian distance covariance," *The annals of applied statistics*, *3*(4), 1236-1265, 2009.