System Health Monitoring of Wind Turbines Using SCADA Data and Gaussian Mixture Models

Akihisa Yasuda[1], Jun Ogata[2], Masahiro Murakawa[3], Hiroyuki Morikawa[4], and Makoto Iida[5]

[1,4,5]*The University of Tokyo Research Center for Advanced Science and Technology, Meguro-Ku, Tokyo, 153-0041, Japan*
yasudaakihisa@mlab.t.u-tokyo.ac.jp
mori@mlab.t.u-tokyo.ac.jp
iida@ilab.eco.rcast.u-tokyo.ac.jp

[2,3]*National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba-Shi, Ibaraki, 305-8560, Japan*
jun.ogata@aist.go.jp
m.murakawa@aist.go.jp

ABSTRACT

Wind turbines are the major driving force to produce renewable energy, but there is a strong need to reduce the costs of operation and maintenance. To detect anomalies of wind turbines, this paper proposes a method which uses data collected by the Supervisory Control And Data Acquisition (SCADA) system and is based upon building the normal behavior model of wind turbines. This is achieved by using supervised data and Gaussian mixture models with filtering SCADA data from the macroscopic point of view. The method is validated with SCADA data collected from actual 2-MW wind turbines. The result shows the potential of detecting anomalies and the effectiveness of filtering conditions for building the model.

1. INTRODUCTION

The Wind power markets has grown strongly worldwide. In 2015, more than 63 GW were added – a 22% increase over the 2014 market – for a global total of around 433 GW and more than half of the world's wind power capacity has been added over the past few years. In order to maintain the growth of these wind power markets, the Levelized Cost of Energy should be kept as low as possible, and the operation and maintenance costs should be reduced. Although an effective method for this purpose is a Condition Monitoring System (CMS), commercially available CMS are very expensive, and attaching to all of wind turbines that exist in every wind farm is very difficult.

Instead of CMS, we propose the use of data collected by the Supervisory Control and Data Acquisition (SCADA) systems, that are already equipped with most of wind turbines. Although the condition monitoring method for wind turbines that utilize the SCADA data is very effective, it is very difficult for previous research to predict the anomaly several months before the failure of the internal equipment of a wind turbine happens, because SCADA data is mainly utilized in the system control of wind turbines and SCADA data is not directly affected by serious damages in the internal equipment.

Therefore, we propose a method to extract the features of wind turbines to build the normal behavior model from SCADA data with proper pre-processing. Further, many previous studies have proposed the method which analyzes individual SCADA data items, but we propose the integration of each filtered SCADA items into one classification model to grasp the dynamic changes of wind turbines sensitively.

2. SYSTEM HEALTH MONITORING MODEL

2.1. System framework for proposed approaches

In this paper, a system framework based on Gaussian Mixture Model (GMM) is proposed for wind turbine health monitoring. This is shown in Figure 1. which includes 2 parts, i.e., off-line modeling and on-line degradation assessment. It needs to collect historical SCADA data from normal behavior state to construct a GMM model and calculate the threshold data. It then uses them to evaluate the performance degradation assessment of wind turbines.

2.2. Strategy of monitoring condition

SCADA data is in the form of time series data and the wind turbines system state is not constant, hence it is very difficult to define the normal behavior model. Therefore, in order to ignore the minor fluctuations and grasp the essential changes

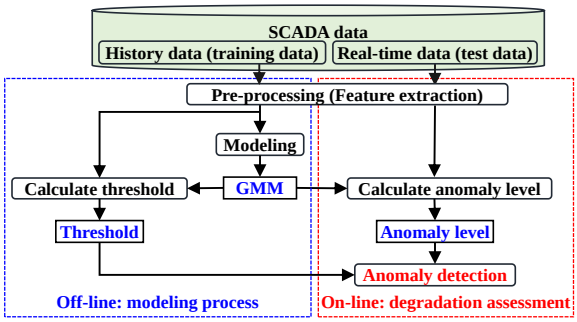

Figure 1. System framework for proposed approaches.

of the internal equipment, we decided to pay attentions to the macroscopic operating condition of wind turbines. Specifically, by constructing a model that uses almost all of SCADA data items related to the internal equipment such as main shaft, gear box, and generator, it is possible to model the entire drive train.

The normal behavior model referred here is the classification model built by historical SCADA data, to classify the current data which is set arbitrarily. The higher the similarity of the normal behavior model to actual wind turbines, the more accurate the evaluation result is.

When building the normal behavior model using machine learning algorithms, it is possible to extract relationships between SCADA data regardless of the specification about structure of the internal equipment of wind turbines.

Further, since SCADA data is time series data, component elements included in the behavior model built from SCADA data are 4. Namely; trend patterns component, cyclic patterns component, seasonal patterns component and irregular patterns component. Because of extracting the characteristics of SCADA data, there is a need to select the most appropriate time series component for building the normal behavior model and to exclude factors of the model referred from other time series elements.

2.3. Characteristics of SCADA data

When making a classification model using machine learning, in order to extract relationships between the data properly, it is important to select the response variable which matched the characteristics of the data. Because SCADA data is used in control applications, wind condition data as input information, power generation data as output information and drive train data as relationship information are collected respectively. Looking at the relationships between these data, it is suitable for SCADA data to build the behavior model of the drive train with respect to the input information by setting the amount of power generation data as output information.

Furthermore, the collection frequency of SCADA data is every 10 minutes and the data used for building the classification model are also mean values. Hence, it is difficult to measure the delicate fluctuations of wind turbines. Therefore, when considering the characteristics of SCADA data in four components of the time series data, the appropriate component for normal behavior model is the trend patterns component. Consequently, when building the normal behavior model using SCADA data, there is a need to extract trend patterns component.

3. PRE-PROCESSING

3.1. Wind conditions

When extracting relationships between SCADA data concerning to power generation, it is necessary to filter SCADA data in wind conditions which are the input information of drive train. Changes in wind conditions are particularly influenced by seasonal fluctuations. Therefore, by performing the filtering of wind conditions, it is possible to simulate the exclusion of the seasonal component.

Firstly, in the filtering process of SCADA data, for the wind conditions, we limit the scope of the wind speed. The wind speed at wind turbines is mainly divided into 3 regions in the operational state. Region 1 is 0 - 5 m/s, Region 2 is 5 - 14 m/s, and Region 3 is 14 - 25 m/s. Focusing on Region 2 is appropriate for health monitoring because it causes moderate load on internal components of wind turbines and induces failure state in the important components such as the gearbox and generator. Therefore, we use only the data that is contained in the wind speed of Region 2 from SCADA data, and exclude all other data.

Secondly, because most of the SCADA data is collected at 10 minutes intervals, we should grasp the wind conditions of the 10 minutes and exclude data which includes strong turbulence factors. Therefore, we use a turbulence intensity I defined as follows, Eq. (1).

$$I = \frac{\sigma^2}{V} \tag{1}$$

In Eq. (1), σ is a standard deviation of wind speed and V is a mean value of wind speed. We determined that the turbulence intensity threshold is 0.5 and all values over this threshold are excluded.

3.2. Operational conditions

There are 2 types of the control method of wind turbines by wind conditions, namely, the pitch control and the yaw control. In the yaw control, we exclude noise data using the number of changes of the yaw angle for 10 minutes. Specifically, we calculate a 95% confidence interval of the number of changes to the yaw angle using the bootstrap method from SCADA data and exclude all the data that crosses the threshold. In the pitch control, we exclude noise data using the average pitch angle for 10 minutes. As well as the number of changes of the yaw angle, we calculate 95% confidence interval of the average pitch angle after excluding

the notable noise data and exclude noise data using threshold of the average pitch angle.

3.3. Noise exclusion

We apply One-Class SVM to exclude the noise data which can't be detected in other pre-processing methods based on domain knowledge about SCADA data. In this One-Class SVM, we set the RBF kernel as a kernel function and as a constraint parameter.

4. CLASSIFICATION MODEL DEVELOPMENT

4.1. Selection of Machine Learning Algorithm

When building a classification model using a machine learning algorithm, classification performance and generalization are the criterion of selecting the algorithm and it is also necessary to consider the characteristics of data. It is very difficult to select the boundary between normal and anomaly by using some supervised-learning model, since there are no prior information regarding the defect severity. In this problem, GMM is the suitable algorithm under the assumption that only the healthy data are available.

GMM model $p(z|\phi)$ with *M* mixture components in R^l can be defined as:

$$p(z|\phi)=\sum_{m=1}^{M}\pi_m p(z|\theta_m) \tag{2}$$

where $z=\{z^{(1)},...,z_l\}^T$ is the *l*-dimensional data vector, $\pi_m \in (0,1)(\forall m=1,2,...,M)$ are the mixing proportions subject to $\sum_{m=1}^{M}\pi_m=1$, θ_m is the parameters composed by the mean vector μ_m and the covariance matrix Σ_m.

In addition, since GMM is a parametric model that extends the Gaussian distribution, if the confidence interval is set to for example 95%, it can be determined that the remaining 5% is an anomaly from the spreading of the probability density distribution. Therefore, even for events with unknown changes such as time series data, thresholds can be set and classified. In the case of non-parametric models such as Neural Net, Support Vector Machine and Random Forest, it is very difficult to set threshold values for anomaly detection, because the parameters of them depend on data completely.

4.2. Negative log likelihood probability

The quantification criterion is needed to evaluate whether a new SCADA record data set is healthy or not. Generally, negative log likelihood probability (NLLP) is used as the quantification indication of health condition, hence, we defined NLLP value as the anomaly score.

For each new SCADA record, NLLP can be described with probability density of GMM as:

$$NLLP=-\log p(z|\phi) \tag{3}$$

In order to improve the sensitivity and reliability of the NLLP to the slight degradation of wind turbine components, NLLP based exponentially weighted moving average (EWMA) statistic is used as an improved quantification indication. EWMA W_t can be obtained as follows:

$$W_t=(1-\alpha)\cdot W_{t-1}+\alpha\cdot NLLP_t \tag{4}$$

Where α is a smoothing constant between 0 and 1, and W_t is obtained by the average of preliminary data. EWMA is a type of convolution, therefore it can be viewed as an example of a low-pass filter and the trend of W_t varies depending on the control parameter α. If the value of α is large, it puts more weight on the current observation than the historic observations. As we denoted in Section 2, we only need to extract the trend patterns component for evaluating the degradation of wind turbines. Hence, in general, a value of α ranging between 0.05 and 0.25 is recommended, but we used 0.01 for proceeding smoothing of fluctuations of operating condition of wind turbines and ignoring cyclic patterns component.

5. EVALUATION

To investigate the effectiveness of the proposed system for wind turbine health monitoring, one actual failure case is studied in this section.

5.1. SCADA data set

SCADA data used in this evaluation is measured in a wind farm which has 10 2 MW-class wind turbines and the data is collected at 10 minutes intervals during the period of 2014/10/01 - 2015/10/31. SCADA data items were selected in relationship to the drive train operation, such as wind speed, rotational speed of main shaft, gear bearing temperature, gear oil temperature, rotational speed of generator, generator bearing temperature, generator temperature and power.

On March 15th 2014, a failure happened in a wind turbine described above. The generator bearing connected with the high speed shaft broke. To evaluate anomaly detection performance of the health monitoring system, we used only SCADA data of a failed wind turbine, because the features relating to the anomaly are averaged and anomaly detection becomes difficult if we use SCADA data of other 9 wind turbines. A period of four-month SCADA data after the failure is selected as the normal condition data. Because the health condition of the wind turbine before the failure has degraded, the health monitoring system would not be able to extract the feature about healthy condition of the wind turbine with that data. Further, we selected SCADA data for 4 month so that the number of records would exceed 2000, and eventually the number of records was 2057.

5.2. Results

Figure 2. shows the NLLP value as an anomaly score which is calculated by the proposed system using only SCADA data. The number of mixture components of a GMM included in the system was 7, and the threshold value was calculated as NLLP value using the training data and indicates a value deviating from the confidence interval 95%. Figure 2. indicates that the health monitoring system has the capability to discriminate between the training data which is the normal behavior state and the test data which includes the anomaly state.

As the evaluation results, we calculated the receiver operating characteristic (ROC) curve using a fixed threshold at the period setting, which is extending every 3 days from the failure occur date. The ROC curve is defined here as the true positive rate (proportion of positive samples that are correctly identified as anomaly with the threshold exceeded) against the false positive rate (proportion of negative samples that are incorrectly identified as anomaly without exceeding the threshold).

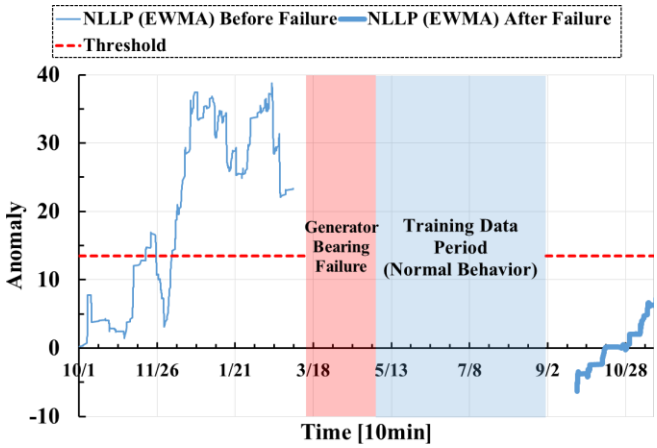

Figure 2. Degradation assessment result by NLLP(EWMA).

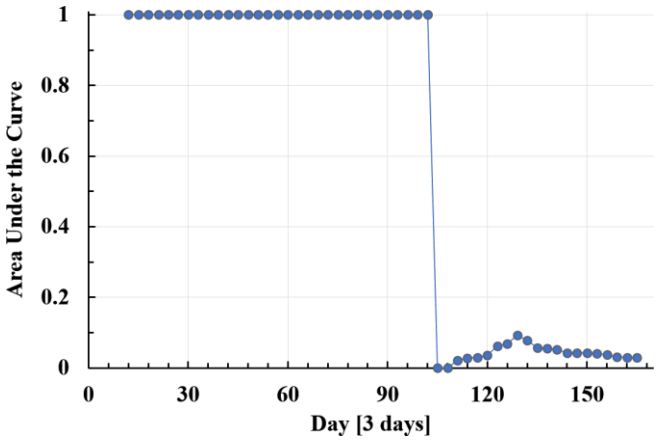

Figure 3. Number of days back from the date of failure.

Figure 3. shows the area under the curve (AUC) which is equal to the area between the full ROC curve. AUC is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Figure 3. indicates the anomaly detection performance of the system and the possibility to predict the failure occurrence three months in advance. In fact, in December 2014 commercially available CMS equipped in a failed wind turbine issued a warning which operators should do maintenance work immediately.

6. CONCLUSION

In this paper, we presented a SCADA data based anomaly detection method for wind turbine health monitoring. At the feature extraction stage, we applied a filtering method based on the domain knowledge about wind turbine operation and selected almost all of SCADA data items related to the internal equipment to model the entire drive train. At the anomaly detection stage, we employed an classification approach based on GMM. The evaluation using actual SCADA data showed that the proposed system has the possibility to predict a failure three months in advance.

ACKNOWLEDGEMENT

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

C. Lins et al. (2015). *Renewables 2015 global status report, REN21*. http://www.ren21.net/renewables-2015-global-status-report-full-report-2

Chris Bishop (2006). *Pattern recognition and machine learning*. Springer.

Duda, Richard O., Peter E. Hart, & David G. Stork (2001). *Pattern classification. 2nd., edition*. New York

Hamilton, James Douglas (1994). *Time series analysis, Vol. 2*. Princeton university press

Rohatgi, J. S., & Nelson, V. (1994). *Wind characteristics: An analysis for the generation of wind power*, Alternative Energy Institute, West Texas A&M University .

Kim, K., Parthasarathy, G., Uluyol, O., Foslien, W., Sheng, S., & Fleming, P. (2011). Use of SCADA data for failure detection in wind turbines. In ASME 2011 5th International Conference on Energy Sustainability (pp. 2071-2079), January, American Society of Mechanical Engineers

Martin, E. B., & A. J. Morris (1996). Non-parametric Confidence Bounds for Process Performance Monitoring Charts. *Journal of Process Control*, 6 6, 349-358

Yu, Jianbo (2011). Bearing Performance Degradation Assessment Using Locality Preserving Projections and Gaussian Mixture Models. *Mechanical Systems and Signal Processing*, 25 7, 2573-2588