Curve-Registration-Based Feature Extraction for Predictive Maintenance of Industrial Equipment

Shouli Zhang[1(✉)], Xiaohong Li[1], Jianwu Wang[2], and Shen Su[3]

[1] Department of School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
zhangshoulia@163.com

[2] Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD 21250, USA

[3] Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, North China University of Technology, Beijing 100144, China

Abstract. With the growing adoption of Internet of Things (IoT), predictive maintenance is gaining momentum for ensuring the reliability of industrial equipment. A common practice of predictive maintenance is to conduct feature extraction on the original sensor data, then conduct deep learning to train predictive maintenance model with the extracted data and finally, conduct prediction by model. Because of the low value density of industrial sensor data stream, feature extraction is usually based on dimensionality reduction. However, traditional methods for dimensionality reduction seldom consider time-lagged correlations which are very common among industrial sensor data streams. More importantly, time-lagged correlations are less sensitive to the traditional dimensionality reduction methods, leading to poor effect of feature extraction. In this paper, we propose a feature extraction method based on curve registration to deal with the time-lagged correlation problem. Our experimental results indicate that our method can: (1) effectively improve the accuracy of prediction; and (2) improve the performance of the prediction model.

Keywords: Predictive maintenance · Time-lagged correlation
Curve registration · Feature extraction

1 Introduction

Predictive maintenance techniques are playing a critical role in a large variety of industries to help anticipate equipment failures to allow for advance scheduling of corrective maintenance. Predictive maintenance is performed based on an assessment of the health status of equipment [1]. Thanks to the rapid development of IoT, massive sensors are deployed on industrial equipment to monitor health status. Thus, a predictive maintenance model can be built by analyzing such sensor data to predict potential failures. With the considerable development of deep learning in recent years, deep-learning-based methods have become popular for predictive maintenance [2–4]. A common practice for predictive maintenance is to conduct deep learning, and feature

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 253–263, 2018.
https://doi.org/10.1007/978-3-030-00916-8_24

extraction plays an important role. This is because the big volume of data streams will result in heavy cost to learning knowledge [2] so that cannot be directly imported into the learning model. And it is difficult to select effective sensors from the massive sensors by their names manually.

In recent years, a number of researchers have applied dimensionality reduction techniques to capture effective features from raw sensor data [5–8]. Feature extraction can be used to transform a set of observations of possibly correlated original data set into a set of values of uncorrelated features by calculating the eigenvectors of the covariance matrix of the original inputs. Thus, extraction effect depends on the correlation between different dimension data. When the correlation of different dimensions is strong, the extraction effect will be obvious and vice versa.

In practical industrial scenarios, there is a common phenomenon that sensor data are correlated with time lags (noted as time-lagged correlation in this paper). On one side, this is from the differences of the system clocks; on the other side, the working principles of the industrial equipment also result in time-lagged correlations among sensor data streams. The time-lagged correlations make the original correlation among sensors seem to be uncorrelated [9], and also decrease the effect of feature extraction. Unfortunately, the existing feature extraction methods [10–13] generally do not consider time-lagged correlation.

In order to improve the extraction effect, we propose a feature extraction method based on curve registration to solve the time-lagged correlation problem. Our challenge here is to figure out the time lags among sensors. To that end, we evaluate the correlations between each pair of sensors based on curve registration methods, and cluster such sensors according to the calculated correlations. With the appropriate time lag adjustment, we manage to map the high-dimensional sensor data to a low-dimension space with adequate original information. The extracted features are passed through an LSTM network, which can be viewed as a deep neural network, to perform predictive maintenance.

2 Related Works

2.1 Feature Extraction Technique for Predictive Maintenance

Feature extraction is one effective way to reduce dimensions and the most import step before training the predictive maintenance model for massive high-dimensional data. Susto [10] employ a supervised regression methodology, called Supervised Aggregative Feature Extraction [11, 12], which exploits a functional learning paradigm to model learning problems with time-series type inputs and scalar output, de facto bypassing the feature extraction phase. Zhang [2] present a log-driven failure prediction system for complex IT systems, which automatically extracts features from IT system logs and enables earlier failure predictions through the LSTM approach on discovering the long-range structure in history data. Kimura [13] adopt a supervised machine learning technique to develop an online template extraction method and a future extraction method that characterizes the abnormality of logs based on the generation patterns of logs. Kimura [14] propose a modeling and event extraction method on network log data using a tensor factorization approach.

Although the above researches are able to automatically extract features from the log files, but their log format is predefined so that it is relatively easy to extract features and template. However, the sensor data is unorganized. Besides, they do not care about the time warping which results in time-lagged correlation among different sensors.

2.2 Predictive Maintenance

Machine-learning approaches are the most popular methods for prognostics. Several quantitative models ranging from simple linear discriminant analysis, more complex logistic regression analysis, and neural networks have been proposed for prediction [15]. Susto develops a multiple classifier machine learning methodology to deal with the unbalanced datasets that arise in maintenance classification problems [16]. But it cannot deal with the high dimensionality of dataset. Baban [17] use a fuzzy logic approach to develop a decision-making system that allows determining the lifetime of the needle and plane predictive maintenance of the needle of a sewing machine. However, it requires expert knowledge and depends on datasets of small quantity. He [18] present an approach for pre-processes sensor signals using short time Fourier transform (STFT). Based on a simple spectrum matrix obtained by an optimized deep learning structure STFT, large memory storage retrieval neural network is built to diagnose the bearing faults. Liu [1] perform a vibration signal analysis to study and extract the behavioral pattern of the bearings. Then use a few machine learning models to classify the type of failure. Finally, they apply a Collaborative Recommendation Approach (CRA) to analyze the similarity of all the model results to suggest in advance.

The above literatures have solved the problem of predictive maintenance to a certain extent, but their method cannot be directly applied on predictive maintenance for large-scale industry because of 'high' and 'wide' sensor data set. In addition, the solution is still mainly dependent on the professional knowledge. However, different industry production environment is various and complex. It is difficult to understand professional knowledge for developers.

3 Problem Analysis

Feature extraction extracts information from the original data as an effective feature [8]. The goal of feature extraction is to find a lower-dimensional data space that will allow projecting the original high-dimensional data on it and form a new presentation of the original data. Herein, we give the formal definition about feature extraction in this paper as following.

Given a sample x_i ($i = 1, \ldots, N$) with H-dimensional vectors lying in a data space S ($S \in R^H$), we need to find a space F ($F \in R^L, L < H$) with L dimension, that makes $x \mapsto x' = F(x)$. In which, x' is the effective features with lower dimension and sufficient information extracted from the original sample data x.

Most feature extraction methods are based on the correlation between sensors. These methods convert the correlated sensor data into a new of linearly uncorrelated presentation (described as feature) by linear transformation. However, in a real production, there is a serious phenomenon that the time warp exists between different sensors which results in time-lagged correlation. This changes the correlation between sensors which can reduce the effect of feature extraction [19].

Next, we explain the time-lagged correlation with a real case in a power plant. There are over ten thousand of sensors deployed on hundreds of equipment in a power plant to monitor machine status at real-time. Figure 1 shows an original sensor data sample generated by a coal mill including 15 sensor data streams. In this paper, we consider Pearson correlation coefficient as the correlation between sensors which is a common used metric. Pearson correlation coefficient between A9 and A10 is 0.171, which indicates that A9 and A10 are irrelevant. But after alignment for A10 with time difference $\Delta t = -4$, Pearson correlation coefficient between A9 and A10 changes to 0.971, leading to the conclusion that A9 and A10 have strong correlation.

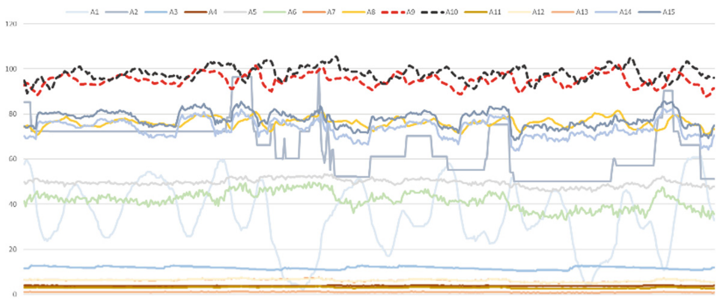

Fig. 1. Example of sensor data stream with time-lagged correlation

To extract feature from the sample data of coal mill, we apply PCA (Principal Component Analysis), which is a common feature extraction method, to the original data set. Without the time difference alignment, the data dimension is reduced from 15 to 7 with totally 98.8% original information. However, with the time difference alignment, the data dimension is reduced from 15 to 6 with the same amount of original information.

According to the above discussion, it is essential to solve time-lagged correlation before feature extraction. In this paper, we focus on the problem of feature extraction with time-lagged correlation and propose a curve-registration-based feature extraction method for predictive maintenance. Time lags for time-lagged correlation sensors are proofread by our method. Finally, we build a predictive maintenance model to predict the failures.

4 Time-Lagged Correlation Oriented Feature Extraction Method

4.1 Curve Registration for Time-Lagged Correlation Sensor Data

Curve registration is an effective way to solve time-lagged correlations. There are many researches on curve registration, the most common one is expectation maximization. The principle of feature extraction is based on correlation between different variables, our curve alignment is based on the correlation maximization. In this paper, we use Pearson correlation coefficient to measure the correlations between sequences. Work [19] presents a correlation-based curve registration method called S-GEM for double sequences. It aims to find a time difference function $\mathrm{d(t)} = d^k(t)$ for two related sequential data TS_1 and TS_2. In which d is the order of function, k is the total number of iterations when computing $\mathrm{d(t)}$. With the help of function $\mathrm{d(t)}$, we can find a $\Delta\mathrm{t} = \mathrm{d}(t_i)$ to make Pearson correlation coefficient $\rho(TS_1(\mathrm{t}), TS_2(\mathrm{t} + \Delta\mathrm{t}))$ maximized.

However, S-GEM is designed for two variables. For a big number of sensors deployed on equipment, such method is not practical if we need to conduct it on each pair of sensors. The first idea came to us is to find a variable as the baseline, then compare it with other remaining variables to find Δt in turn. However, this method brings up two problems: (1) which one should be selected as the baseline; (2) what should be done with the irrelevant variables compared to the baseline variable.

As a further thought, we cluster all variables into groups, and conduct curve registration within each pair of variables in the same group. To make the choice of the baseline variable that is not sensitive to the clustering result, we only cluster two variables into the same group when they are strongly correlated. And we manage to consider no curve registration problem in this way. It is the fundamental assumption of our method that sequences in the same cluster have the same correlation. Thus, every sequence in the same clusters can be selected as a reference curve, and we select the first one as curve registration. Other sequence can be aligned by S-GEM according with reference curve in turn.

Formally, our method works as following. Let $\{X_1, X_2, \ldots, X_h\}$ be h multivariate curves that can be considered as a waveform of observers for multi sensor data. Given at iteration $v = 1$, a K-means based clustering algorithm is applied on the original sensor data to cluster the wave curve into k groups $\mathrm{P}^1 = \{G_1^1, G_2^1, \ldots, G_k^1\}$, and P refers to the partition. For every group, we select a random sequence as their reference curves as to form k references. Combined with S-GEM curve registration algorithm, we compared the other sequences with reference curves.

The procedure of alignment for sequences with time-lagged correlation is as following:

(1) *At the v-th iteration, compute the reference curves $R_1^v, R_2^v, \ldots, R_k^v$ for each group;*

(2) *For each curve X_i^v, i = 1, ..., d, compute the maximal Pearson correlation coefficient with the reference curve R_m^v of each group G_m^v of the current partition by S-GEM and get a Δt to compute the aligned curve $X_i'^v = X_i^v(t + \Delta t)$.*

(3) *Assign each curve* $X_i'^v$ *to the group* G_i^* *which has the maximal Pearson correlation coefficient, then update the partition accordingly:* $P^{v+1} = \{G_1^{v+1}, G_2^{v+1}, \ldots, G_k^{v+1}\}$;

(4) *Update v and repeat from step (1) until no movement of elements among groups is observed in step (3).*

This method, structured as the k-means method, deals with the simultaneous optimal clustering and warping of curves, implemented by alternating Pearson correlation maximization steps. Every curve of sensor data has a maximize Pearson correlation coefficient with each other in the same group without time-lagged correlation.

4.2 Feature Extraction Using PCA

Principal component analysis (PCA) is a well-known feature extraction method. The basic principle of PCA is to collate the information from the interrelations of the variables. Then, PCA uses the orthogonal transforms to find the variable group with high correlation and realize the compression of the observed data to a lower dimension. After transformation with PCA, the greatest variance by any projection of the data comes to lie on the first principal component, the second greatest variance on the second coordinate, and so on.

Given a set of N observations, which is constructed with p-dimensional random variable $X = X_1, X_2, X_3 \ldots X_p$. The principle of PCA is to compute a new matrix $Y = A^T X$. $A = (A_1, A_2, \ldots A_m)$ is an orthogonal matrix.

The process of extracting the features from X is to find out A. The procedure works as following:

(1) *Normalization the data set with empirical mean;*
(2) *Calculate the covariance matrix for multi variables;*
(3) *Calculate the eigenvalue and eigenvectors for the covariance matrix;*
(4) *Sort the eigenvalues according to the contribution of each principal component. Select the first m eigenvalues and eigenvectors according an information threshold (noted as contribution)* φ *to form A.*
(5) *Calculate the final principal component Y by the transformation* $Y = A^T X$.

4.3 Predictive Maintenance Model

In this paper we take the failure prediction as a binary classification problem. By giving the input sequence and their labels, the output of prediction model is the probability of a forthcoming failure. The latter status of equipment depend on a long historical trend. Long Short Time Memory (LSTM) is a kind of Recurrent Neural Network (RNN), which can be viewed as a deep neural network. It is designed to improve storing and accessing information compared to classical RNNs. LSTM has recently been successfully applied in a variety of sequence modeling tasks [3]. Thus, we apply a LSTM-based network to build the predictive maintenance model.

The input feature vector sequence $x = (x_{t-L+1}, \ldots, x_t)$ is passed to a stack of multiple recurrently connected hidden layers through weighted connections to train the model. And the output is the sequence $y = (y_{t-L+1}, \ldots, y_t)$. The output y_t is a binary vector serving as a representation of the system status which can be used to parameterize the probability distributing $P\ (d_t|y_t)$ of the target d_t. The target d_t is a binary vector with 2 complementary classes.

$$\mathrm{P}(d_t = \mathrm{k}|y_t) = \hat{y}_t^k = \frac{\exp\left(y_t^k\right)}{\sum_{k'=1}^{K} \exp\left(y_t^{k'}\right)}$$

where k is the number of classes and therefore, in our case $k = 2$. Given a threshold value $\mathcal{E}$, if the P is greater than $\mathcal{E}$, the predictive value can be regarded as an anomaly signal. On the contrast, it is a normal signal.

5 Experiments and Evaluation

5.1 Data Set and Environment

The experimental evaluation of our method is conducted on a cluster of 8 nodes, each node with 8-core Intel Xeon (E312xx) 8 GB processors and 32 GB of RAM, interconnected with 1 GBs Ethernet and each run in virtual machines with CentOS 6.4 and java 1.8.

The datasets used in our experiment is the real sensor data collected from a coal-fired power plant. We collect our experimental data from "primary air fan", "secondary air fan", "coal mill", "force draft fan" and "induced draft fan". The sensor number of them are respectively 45, 37, 49, 38 and 42. And the data recorder is 278400. The failure events in failure logs are regarded as some baselines to verify the predictive results.

5.2 Evaluation Metrics

For our evaluations, we consider the following performance metrics related to our prediction problem.

Precision: precision represents how many abnormities are accurate according to failure logs, which can be defined as following:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall: Recall is a metric of the proportion of all possible correct results that our method actually discovers, which can be defined as following:

$$\text{Recall} = \frac{\text{True Positive}}{\text{All Positive}}$$

Contribution: Contribution presents the amount of information remained in extracted features compared to original data.

Dimension: Dimension means the dimension after dimension reduction of extracted features.

5.3 Experiment and Evaluation

Our predictive model aims at helping finding abnormity in advance, and has deployed in the real production. We compare our method with the following method to detect the abnormalities in Power Plant:

rule-based: Traditional rule-based methods to detect the abnormalities based on accumulative experience.

p-LSTM: We firstly extract features from original sensor data using PCA, the enter the features into LSTM neuro network to predict early failures.

dp-LSTM: Our proposed method, which firstly preprocess the original sensor data by cure registration, then extract features from curve registrated sensor data using PCA, finally enter the features into LSTM to predict early failures.

First, we conduct the feature extraction and compare the **Dimension** under the same contribution of 98.9%. The result is shown in Table 1. The average dimension reduction with dp-LSTM is 28.86% of the original dimensions, and the average dimension reduction with p-LSTM is 38.37% of the original dimensions. Table 2 shows the result of Contribution under the same dimension.

Table 1. Dimension reduction with same contribution

Device	Sensors	p-LSTM	dp-LSTM	Contribution
Primary air fan	45	20	16	98.9%
Secondary air fan	37	17	11	98.9%
Coal mill	49	13	10	98.9%
Force draft fan	38	14	11	98.9%
Induced draft fan	42	16	12	98.9%

Table 2. Contribution with same dimension

Device	Sensors	Extracted features	p-LSTM	dp-LSTM
Primary air fan	45	16	97.8%	99.51%
Secondary air fan	37	11	96.45%	99.62%
Coal mill	49	10	97.3%	98.71%
Force draft fan	38	11	97.76%	99.18%
Induced draft fan	42	12	97.8%	99.18%

Then we build a LSTM network with 4 hidden layers and 50 hidden units in each layer. We initialize all weight parameters uniformly in the range [−0.08, 0.08], while initializing the LSTM forget gate with a slightly higher bias (set bias value to 1.0) to encourage remembering at the beginning. We then train the network using mini-batch stochastic gradient descent with learning rate 0.001 and decay factor 0.95. We train each model for 50 epochs and decay the base learning rate after 10 epochs 6 by multiplying it with the decay factor 0.95 for each additional epoch. We select the first 80% as training data and the rest as testing data.

Figure 2 shows the precision and recall for different equipment with different methods. The average precision of rule-based method is 0.532. The average precision of p-LSTM is 0.714. The average precision of dp-LSTM is 0.78. The average recall of rule-based method is 0.474. While the average recall of *p-LSTM* is 0.78. The average recall of *dp-LSTM* is 0.856. It has virtually objectively proved that the deep learning-based predictive maintenance is more effective than the traditional rule-based method. Benefit from our curve registration for time-lagged correlation sensor data, it can be concluded that our method significantly contributes to the performance of the feature extraction and model applied.

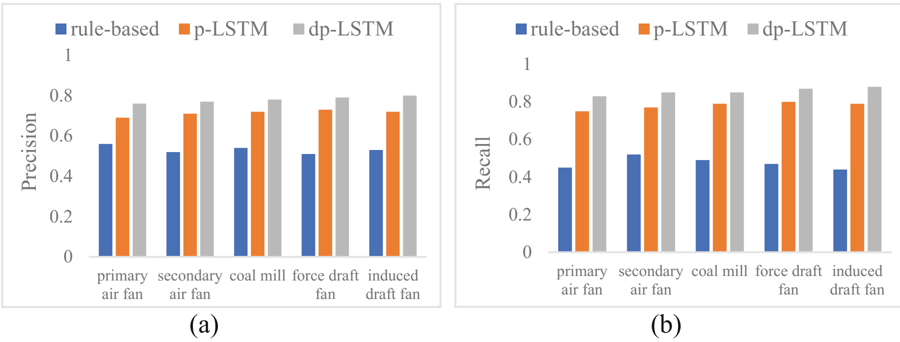

Fig. 2. Precision of different method for early failures prediction

6 Conclusion

The high-dimensional data for predictive maintenance and time-lagged correlation among sensors result in challenges to build predictive maintenance model. In order to extract effective features from massive high-dimensional sensor data with time-lagged correlation, this paper presents a time-lagged correlation based feature extraction method for predictive maintenance model. According to the experiments, we prove that: (1) our feature extraction method manages to improve the dimension reduction compared to the traditional feature extraction method while ensuring the contained original information; (2) our extracted features manage to reduce the training time of predictive maintenance resolution with assurance of predication accuracy.

Acknowledgments. This work was supported in part by a grant from the National Natural Science Foundation of China (Grant No. 61672042), the Program for Youth Backbone Individual, supported by Beijing Municipal Party Committee Organization Department (Grant No. 2015000020124G024).

References

1. Liu, Y.: Predictive modeling for intelligent maintenance in complex semiconductor manufacturing processes (2008)
2. Zhang, K., Xu, J., Min, M.R., et al.: Automated IT system failure prediction: a deep learning approach. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 1291–1300. IEEE (2016)
3. Sipos, R., Fradkin, D., Moerchen, F., et al.: Log-based predictive maintenance. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1867–1876. ACM (2014)
4. Rawat, S., Rawat, S.: Multi-sensor Data Fusion by a Hybrid Methodology—A Comparative Study. Elsevier, Amsterdam (2016)
5. Qiu, X., Fu, D., Fu, Z.: An efficient dimensionality reduction approach for small-sample size and high-dimensional data modeling. J. Comput. **9**(3), 576–580 (2014)
6. Sharma, N., Saroha, K.: A novel dimensionality reduction method for cancer dataset using PCA and feature ranking. In: International Conference on Advances in Computing, Communications and Informatics, pp. 2261–2264. IEEE (2015)
7. Abualigah, L.M., Khader, A.T., Al-Betar, M.A., et al.: Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. Expert Syst. Appl. **84**, 24–36 (2017)
8. Chang, P.C., Wu, J.L.: A critical feature extraction by kernel PCA in stock trading model. Soft Comput. Fusion Found. Methodol. Appl. **19**(5), 1393–1408 (2015)
9. Adelfio, G., Chiodi, M., Luzio, D., et al.: Simultaneous seismic wave clustering and registration. Comput. Geosci. **44**(13), 60–69 (2012)
10. Susto, G.A., Beghi, A.: Dealing with time-series data in predictive maintenance problems. In: 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–4. IEEE (2016)
11. Schirru, A., Susto, G.A., Pampuri, S., McLoone, S.: Learning from time series: supervised aggregative feature extraction. In: 2012 IEEE 51st Annual Conference on Decision and Control (CDC), pp. 5254–5259. IEEE (2012)
12. Susto, G.A., Schirru, A., Pampuri, S., McLoone, S.: Supervised aggregative feature extraction for big data time series regression. IEEE Trans. Industr. Inf. **12**(3), 1243–1252 (2016)
13. Kimura, T., Watanabe, A., Toyono, T., et al.: Proactive failure detection learning generation patterns of large-scale network logs. In: International Conference on Network and Service Management, pp. 8–14. IEEE Computer Society (2015)
14. Kimura, T., Ishibashi, K., Mori, T., et al.: Spatio-temporal factorization of log data for understanding network events. In: 2014 Proceedings IEEE INFOCOM, pp. 610–618. IEEE (2014)
15. Warriach, E.U., Tei, K.: Fault detection in wireless sensor networks: a machine learning approach. In: 16th International Conference on Computational Science and Engineering, pp. 758–765 (2013)

16. Susto, G.A., Schirru, A., Pampuri, S., et al.: Machine learning for predictive maintenance: a multiple classifier approach. IEEE Trans. Industr. Inf. **11**(3), 812–820 (2015)
17. Baban, C.F., Baban, M., Suteu, M.D.: Using a fuzzy logic approach for the predictive maintenance of textile machines. J. Intell. Fuzzy Syst. **30**(2), 999–1006 (2016)
18. He, M., He, D.: Deep learning based approach for bearing fault diagnosis. IEEE Trans. Ind. Appl. **53**(3), 3057–3065 (2017)
19. Jiang, G.X., Wang, W.J.: Correlation analysis in curve registration of time series. J. Softw. **9**, 2002–2017 (2014)