Multivariate time series classification for equipment monitor and fault detection

Wei-Chen Liu
Department of Information Management
Yuan Ze University, Taoyuan, Taiwan
Email: dorgonway@gmail.com

Chia-Yu Hsu†
Department of Information Management
Yuan Ze University, Taoyuan, Taiwan
Tel: (+886) 3- 4638800 ext. 2793, Email: cyhsu@saturn.yzu.edu.tw

Abstract. Through extracting more useful information from the sensor data, temporal analysis and time series data analysis have received more attentions and applications. First, the discriminative feature was extracted from the original time series data, called Shapelet, and Symbolic Aggregate approXimation-Vector Space Model (SAX-VSM) was used to build model for fault detection. Second, the similarity between the time series were directly calculated for difference detection by 1 nearest neighbor classification with Euclidean distance (1NN-EUC) and Dynamic Time Warping(1NN-DTW). However, most of the existing studies mainly focus on univariate time series and are difficult to solve the multivariate time series problem. This paper proposes a multivariate time series model for fault detection from the streams of sensor data. We also conduct an empirical study to demonstrate that the proposed approach outperform than other existing time series analysis model.

Keywords: time series data, stream data, fault detection, equipment monitor, big data

1. INTRODUCTION

With the rise of Industry 4.0 and Internet of Things (IoT), large volume streaming sensor data are accumulated quickly. There are many sensors collect the data from the machine with the advanced information and sensor technology. The maintenance for the expensive machine is become seriously. A large number of data for each sensors are continuously collected by secondly such as pressure, temperature. Through capture the equipment abnormality by detecting change of time series profile from these stream of sensors, it is crucial to predict the equipment health and increase yield (Lee et al. 2013). The multivariate time series from sensors is used to identify the key feature and extract of useful information through insight into the sensor becomes importantly (Patri et al., 2014). The information can be regarded as fault detection and find the difference among the sensor profile to assist the engineer abnormality diagnosis quickly (Chien et al. 2013).

Time series analysis has been applied in different fields such as speech recognition (Sakoe and Chiba, 1978), health care (Kampouraki et al., 2009), Fault diagnosis (Chen and Feng, 2013) and sensor network (Patri et al., 2014). To measure the similarity between time series, 1-NN classifier with Euclidean (Faloutsos et al., 1994) and DTW (Sakoe and Chiba, 1978) were used. SAX-VSM (Senin and Malinchik, 2013) and Shapelet (Ye and Keogh, 2009; Rakthanmanon and Keogh, 2013) were integrated to extract the discriminative subsequence from time series data. However, most of the existing studies mainly focus on univariate time series and are difficult to solve the multivariate time series for the equipment monitoring and fault detection.

This paper aims to propose a framework can detect failure from the multivariate time series, in which Shapelet is used for feature extraction and then random forest is applied to conduct a model for fault detection. To demonstrate the proposed method, we also conduct an empirical study from a fabrication. The experiment results show that our approach outperform other method for detecting equipment abnormality.

The rest of paper is organized as follows. A briefly discuss on related work in Section 2. Section 3 illustrates our approach and an empirical study was conducted in Section 4. Finally, we offer conclusion and suggestions for future work in Section 5.

2. RELATED WORK

There are primary two approaches in time series analysis. First, extracting the important feature from the time series, such as Shapelet (Ye and Keogh, 2009; Rakthanmanon and Keogh, 2013) and Symbolic Aggregate approXimation-Vector Space Model (SAX-VSM) (Senin and Malinchik, 2013). The another type of algorithms is to calculate the similarity by 1 nearest neighbor classification with Euclidean distance (1NN-EUC) (Faloutsos et al.,1994) and Dynamic Time Warping (1NN-DTW) (Sakoe and Chiba, 1978). Their advantages are simple and accurate without a training model. The definitions and notations used in this paper are defined as follows:

Definition 1: A subsequence S_k of T length $m \leq n$ is a contiguous sequence from T with start position at k, $S_k = t_k, \dots, t_{k+m-1}$ for $1 \leq k \leq n - m + 1$.

Definition 2: The sliding window is sliding a window size n across T and getting all of subsequence S_k. The number of the subsequence is $m = n - m + 1$.

Definition 3: The distance between two the same of length m series. Frist, we perform Z-normalization to the time series before calculate Euclidean Distance. Assume x is shapelet and y is subsequence from sensor are the same length, the distance is defined as (1):

$$\text{dist}(x, y) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2} \qquad (1)$$

Definition 4: The minimum distance is calculated the minimum distance between the same length subsequence S and any subsequence of T.

The Shapelet is an algorithm for find the most discriminative subsequence in time series (Ye and Keogh, 2009). The shapelet also called "snippet", it can create a model to classify the observation. This approach searches all of the possible result to find the best shapelet with high classification accuracy. Rakthanmanon and Keogh (2013) provided the effective algorithm to find the shapelet called Fast Shapelet, which is incorporated by Symbolic Aggregate approXimation (SAX) to find the shapelet. Although Fast Shapelet doesn't find the best shapelet, however, the computation efficiency can be improved.

Patri et al. (2014) use shapelet to solve the multivariate problem time series called Shapelet Forest. Shapelet Forest uses Fast Shapelet to find the shapelet form each univariate time series and uses the shapelet to calculate the distance with the subsequence from corresponding time series, then learns weight for each distance and voting the final result.

Senin and Malinchik (2013) proposed Symbolic Aggregate approXimation-Vector Space Model (SAX-VSM) to find the discriminative subsequence from the time series. Symbolic Aggregate approXimation (SAX)(Faloutsos et al., 1994) is the method use Piecewise Aggregate Approximation (PAA) (Keogh and Pazzani, 2000) to transform time series to SAX word. Vector Space Model (VSM) (Salton et al., 1975) is known in information retrieval for calculating the cosine similarity between vectors. SAX-VSM use SAX to get the subsequence with time series then do the VSM to find the most similarity subsequence.

However, most of the approaches mainly focused on univariate time series. To bridge the gap between the existing studies, this study propose an approach which is capable to simultaneously handle the multivariate time series for equipment monitoring and fault detection.

3. APPROACH

This section describes our approach framework as the figure 1. First, the data preparation is shown. Second, we describe how to find the shapelet from each of time-series sensor data. Next, we calculate the minimum distance between the shapelet and the sensor's subsequence. Eventually, random forest algorithm is used to train our model and predict the equipment status either in normal or abnormal.

Figure 1: Framework of approach

3.1 Data Preparation

Due to there are many sensors used to collect the data when processing the wafer, each of sensor has their own different scale. In data preparation, we do z-normalization to the all of the sensor, ensure all of the univariate time series data in the same scale range like as shown in Figure 2. The definition of z-normalization is given in (2).

$$x' = \frac{x - \mu}{\sigma} \qquad (2)$$

where x is the value of point in time series, μ is the mean of time series, σ is the standard deviation of the time series and x' is the normalized value.

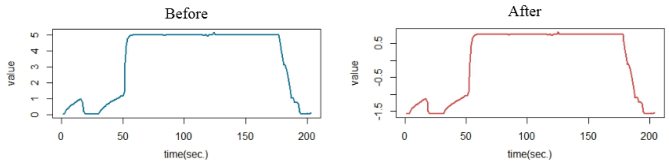

Figure 2: Example of z-normalization

3.2 Shapelet Extraction

The Fast Shapelet algorithm is used to effectively extract the shapelet from each sensor as the feature. Because the wafer is monitored by various sensors, and then each corresponded sensor is labelled the same as the wafer. For example, the second wafer is normal then second wafer's entire sensors are labeled as normal. It may cause a problem about the sensor is abnormal but label it is normal. Despite this situation, we will show that our approach can deal with this problem in final result.

The shapelets are our features for the wafer abnormality and represent the discriminative part of each sensor. These shapelets are used to classify whether the single sensor is normal or abnormal. For example, there are three shapelets from sensor 1. Figure 3 shows the example about shapelet, red one is the shapelet from sensor 1 and sensor 14.

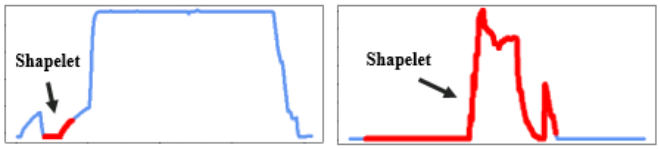

Figure 3: Shapelet from the sensor1 (left) and sensor14 (right)

3.3 Distance Matrix

A data matrix as $D = \{d_{ij}\}$ is created for each sensor to search their shapelets and use them for distance calculation between the shapelet and the subsequence of sensor. Then, we use sliding window through the sensor get subsequence of sensor according to Definitions 1 and 2. We make a distance matrix to fill in calculating Euclidean distance between each of shapelet and corresponded sensor by Definition 3. For example, d_{13}, the shaplet 1 is assumed to extract from sensor 3 through slide window, and then we get subsequences from sensor 3. Therefore, Euclidean distance between subsequences of sensor 3 and shapele 1 is computed and only the minimum distance is selected to regard as the new feature by Definition 4.

3.4 Prediction

Random forest select bootstrap sample from the train data, then create number of n tree to classify the data. The final result is decided by voting on each of tree. Figure 4 presents the architecture of random forest. According to the previous distance matrix to train model. The variables are the minimum Euclidean distance between shapelets of each sensor and the subsequence of the correspond sensor.

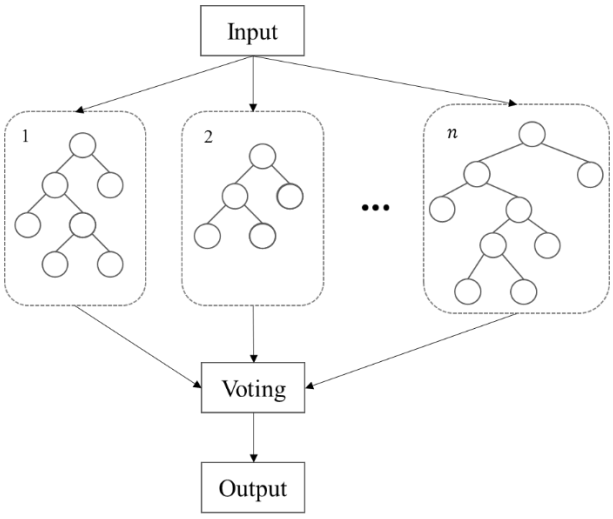

Figure 4: The architecture of random forest

4. EXPRIMENT

We propose an approach for time series classification based on Shapelet and compare the classification performance among decision tree, random forest, support vector machine and neural network, which is capable of identifying which part of equipment sensor caused wafer abnormality.

4.1 Dataset

The dataset is the sensor data collecting during the wafer processing with 21 sensors. There are 147 normal wafers and 41 abnormal wafers. Through eliminating the sensor with constant value, total 17 sensors are selected as input variables for model construction. Three different sensors about the dataset are shown in Figure 5.

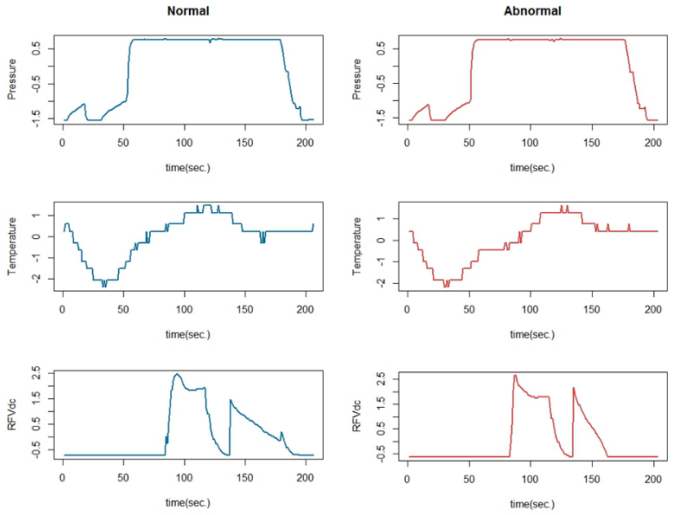

Figure 5: Sample of collected sensor

4.2 Result and Discussion

To evaluate the performance of our approach, five-folds cross validation is used to detect the abnormal wafer. Table 1 showed that the random forest algorithm has the best performance, so we choose the random forest and set the number of tree is 10,000 to get the importance (mean decrease Gini) of each shapelet from each sensor as shown in Figure 6. Those shapelets are extracted by each sensor. For example, assuming that the sensor 1 has four shapelets, and then we select the importance by 5-folds-cross validation and select the highest one to link with the sensor1. Figure 6 shows that the sensor 14 is the most important part in all of them, which indicates the sensor 14 maybe could be high correlated to the abnormality.

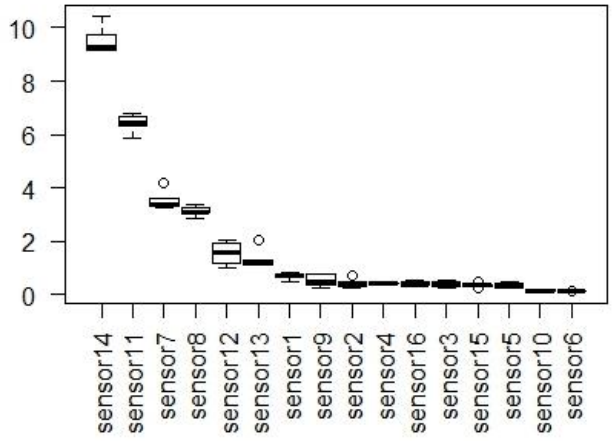

Figure 6: Mean decrease Gini of each sensor

We choose the sensor14 to show the distance between sensor14 and shapelets from the sensor 14 in Figure 7. Note that there are only one shapelet extracted from sensor14. Figure 7 shows that the difference in distance between subsequences of sensor which is labelled as normal or abnormal and the shapelet from the sensor14. Through the distance between shapelet and sensor profile, we can identify the wafer in normal or abnormal.

To evaluate the model performance, precision and recall, which are two classification-oriented measures, are defined as follows:

$$\text{precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

where True Positive (TP) is the number of abnormal condition that are correctly classified into abnormality, False Positive (FP) is the number of abnormal condition that are incorrectly classified, and False Negative (FN) is the number of normal condition that should be classified, but not be determined incorrectly.

Tables 2 to 4 summarize the results of precision, recall, F-score among random forest, decision tree, support vector machine, and backpropagation neural network. Random forest is superior to other methods. In particular, support vector machine has high precision but the recall is low. It shows that our model can extract the discriminative feature and has insight into the sensors.

Table 1: Classification summary of five-folds cross validation

	precision	recall	F-score	AUC (%)
random forest	**1.000**	**0.897**	**0.940**	**97.86**
decision tree	0.897	0.879	0.875	94.68
support vector machine	0.980	0.785	0.857	94.68
neural network	0.875	0.855	0.855	93.64

Table 2: 5-fold cross validation: precision

method	5-Folds cross-validation				
	1	2	3	4	5
random forest	1.00	1.00	1.00	1.00	1.00
decision tree	1.00	0.78	0.89	1.00	0.82
support vector machine	1.00	1.00	1.00	1.00	0.90
neural network	1.00	0.56	1.00	1.00	0.82

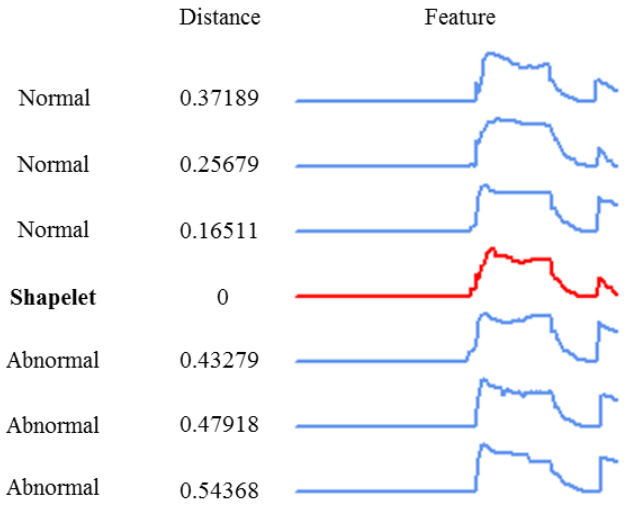

Figure 7: Shapelet and subsequence of sensor14

Table 3: 5-fold cross validation: recall

method	5-Folds cross-validation				
	1	2	3	4	5
random forest	1.00	1.00	0.818	0.667	1.00
decision tree	1.00	1.00	0.727	0.667	1.00
support vector machine	1.00	0.714	0.545	0.667	1.00
neural network	1.00	0.714	0.727	0.833	1.00

Table 4: 5-fold cross validation: F-score

method	5-Folds cross-validation				
	1	2	3	4	5
random forest	1.00	1.00	0.90	0.80	1.00
decision tree	1.00	0.875	0.80	0.80	0.90
support vector machine	1.00	0.833	0.706	0.80	0.947
neural network	1.00	0.625	0.842	0.909	0.90

5. CONCLUSION AND FUTURE WORK

This paper proposes a multivariate time series model with shapelet extraction from the streams of sensor data for fault detection. We incorporate the whole sensor profile instead of considering part of key step. It can identify the difference from the amount of large sensor profile and provide the information for fault diagnosis. According to the empirical study, we demonstrate that the proposed approach outperform than other existing time series time model. According to the analysis results, the shapelet with random forest algorithm outperform the other classification methods. The further research, we can apply the proposed framework into other engineering domain regarding the fault detection.

ACKNOWLEDGMENTS

This study was supported by Ministry of Science and Technology, Taiwan (MOST 103-2221-E-155-029-MY2).

REFERENCES

Chen, Z. and W. Feng (2013). Detecting impolite crawler by using time series analysis. *The 25th International Conference on Tools with Artificial Intelligence*, 123-126

Chien, C.-F., Hsu, C.-Y., and Chen, P. (2013). Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence. *Flexible Services and Manufacturing Journal*, 25: 367-388.

Faloutsos, C., Ranganathan, M., and Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. *Proceedings of the ACM SIGMOD international conference on Management of data,* Minneapolis, 419-429.

Kampouraki, A., Manis, G., A Nikou, C. (2009). Heartbeat time series classification with support vector machines. *IEEE Transactions on Information Technology in Biomedicine*, 13(4): 512-518.

Keogh, E. J., and Pazzani, M. J. (2000). A simple dimensionality reduction technique for fast similarity search in large time series databases. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 122-133.

Lee, J., Lapira, E., Bagheri, B., and Kao, H. A. (2013). Recent advances and trends in predictive manufacturing systems in big data environment. Manufacturing Letters, 1(1): 38-41.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2-11.

O. P. Patri, A. Panangadan, C. Chelmis, R. G. McKee and V. K. Prasanna (2014). Predicting failures from oilfield sensor data using time series shapelets. *SPE 170680-MS presented at the SPE Annual Technical Conference and Exhibition*.

Patri, O. P., Sharma, A. B., Chen, H., Jiang, G., Panangadan, A. V., and Prasanna, V. K. (2014). Extracting discriminative shapelets from heterogeneous sensor data, *2014 IEEE International Conference on Big Data*, 1095-1104

Rakthanmanon, T., and Keogh, E. (2013). Fast shapelets: A scalable algorithm for discovering time series shapelets. *In Proceedings of the 13th SIAM International Conference on Data Mining*, 668-676.

Sakoe, H. and S. Chiba (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1): 43-49.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM,* 18(11): 613-620.

Senin, P. and S. Malinchik (2013). Sax-vsm: Interpretable time

series classification using sax and vector space model. *In Proceedings of the IEEE 13th International Conference on Data Mining*, 1175-1180.

Ye, L. and E. Keogh (2009). Time series shapelets: a new primitive for data mining. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 947-956.