THE PROBABILITY PREDICTION OF FAULT STATE USING CASE-BASED REASONING WITH RANDOM FOREST

[1]BONG-GEUN SONG, [2]SEUNG HWAN PARK, [3]CHEONG-SOOL PARK, [4]JUN-GEOL BAEK

[1,2,3,4]Department of Industrial Management Engineering, Korea University
E-mail: [1]sbk0809@korea.ac.kr, [2]udongpang@korea.ac.kr, [3]dumm97@korea.ac.kr, [4]jungeol@korea.ac.kr

Abstract— Semiconductor process is that have spent a lot of economic and time costs when an abnormality occurs. It is essential to manage the efficient process monitoring about the abnormality diagnosis of the equipment. Since the vacuum pump are used in a number of steps, it is possible to diagnose the overall condition of the process through the health monitoring via vacuum pump. In this paper, we performed an analysis using the case information to be generated in response to the aging of equipment in case based reasoning (CBM). In order to optimize the case which is defined as Initial state, Aging state, Fault state defined, we determine the two experimental conditions and analysis using a machine learning technique. In analysis, we used the Random Forest that there is strength in multiple and imbalance class. Based on the optimal case that selected as the experimental results, we can provide decision to assist with the diagnosis and prediction of equipment.

Keywords— Vacuum pump, Random Forest, case-based reasoning, Initial state, Aging state, Fault state.

I. INTRODUCTION

It is necessary to quickly and accurately measure the abnormality of the process in order to efficiently manage the process.Fault, Degradation, Damage that occur during the process operations will be induced decrease in yield such as production time, decreased quality, economic loss. Accordingly, by predicting the status of the equipment occurring in the process system in advance, we can reduce costs and make optimal management plans. Therefore,it is essential to develop a Logic which can monitoring the various forms of data obtained from the facility. If we properly handle vast amounts of data, it is possible to obtain the improved productivity and early detection of abnormality.

It is proceeded to hundreds of processes to make one semiconductor product and can be devided into four major steps. There is a FAB(fabrication), Assembly process, Package test process. In semiconductor process, such as chemical vapor deposition(CVD), vacuum pump, Iron implantation are mainly needed to the process of requiring a vacuum environment(Lessard et al, 2000). Vacuum pump used to form a vacuum, so more chemicals are uniformly applied to the wafer surface in the chamber. Also since the role of removing the gas and impurities, vacuum pump significantly impacts on yield. For these reasons, It is important that the vacuum and impurity are removed and the irons in the chamber moves freely through the vacuum pump.

Until now, we have used the data generated by the etching and step in the chamber. However, the research to diagnose the abnormal equipment state using the pump data is insufficient. In the characteristics of a semiconductor process the use of hazardous and corrosive chemical gas so frequently, the vacuum pump is used for outflow these materials and formation of a vacuum. Vacuum pump plays an important role in creating an enviroment in process chamber in a semiconductor manufacturing process, unless there is a specific reason, continuous operation. As current vacuum pump are used in equipment of various fields such as display and semiconductors, the research to diagnose the abnormality in the data generated from the pump has the following advantages. (1)The pump maintains the state of the process to be uniform and removes impurities. (2) It is not possible to predict the state of the pump which plays an important role in the process when the data in the chamber occurs. (3) While the chamber determines a wafer yield, pumps have been associated in a number of wafer yield. So it is possible to efficiently fault management.

In this paper, we used the sensor data generated by a diffusion vacuum pump of semiconductor process and carried out research to understand the health condition of equipment. Existing research on diagnosis and prediction of equipment extracts the pattern characteristic of monotonic increase or decrease from the signal and made the health indicators.

(Mosallam et al, 2013) uses the data of the bearing provided by the NASA repository. They extracts the trend information using Empirical Mode Decomposition (EMD) and predicts the Remaining Useful Life(RUL) in ridge regression. (Coble et al, 2009) utilizes Auto-Associative Kernel Regression (AAKR) to the signal to extract the residual. And then they proposed a General Path Model (GPM) to predict the RUL by using the features that satisfy the Monotonicity, Prognosability, Tradeability defined in

the paper. As this same model, (coble et al, 2011)applies a Bayesian method. It was shown to predict effectively RUL in data of a small size by use of extrapolation. (Barbieri et al, 2015) extracts the feature in time domain and frequency domain and applies to GPM. In general, they predicts the RUL by using the parameter that better represents the trend of the signal. However, it is difficult to apply the existing methods when increase and decrease pattern is not extracted well from data. The Data used in this study, although each of the sensor data moves organically e.g. when the pressure and the current increases the temperature decreases, is difficult to see that they represent a pattern of monotonically increasing or decreasing.

In this paper, we conducted to approach as a classification problem using the health state by defining the inherent state in the equipment. In previous research, (Ali, et al. 2015) analysed the fault diagnosis based on vibration signals in bearing.It was analyzed using EMD and artificial neural networks (ANN). Seven classes are defined to follow bearing degradation over time. And to detectearly stage bearing degradations in real time, this work proposes a health index. The paper has a failure information and the bearing forms a similar pattern in accordance with the type of fault. However, the data of this research, there is no information of fault type and change in the signal pattern which is significantly displayed in response to a bearing fault type. Therefore, in order to solve the above problems, perform the following experimental procedure. First, define the status of Initial, Aging, Fault in order to understand the inherent state in the equipment. For optimization the above states we conducted the following two experimental conditions.(1) Optimum length of above states(2)Model comparison of 2-Class(Initial+Aging, Fault) versus 3-Class(Initial, Aging, Fault).Finally, in order to understand the status of the equipment in real time, providing an indication to the health monitoring by using the Fault state information derived from the experimental results.

II. APPROACH

2.1. Case-Based Reasoning(CBR)

In general, the failure of the process have been resolved through the corrective maintenance and preventive maintenance. However, because there is a limit in terms of cost and reliability, the new technology has been demanded. In this study, in order to solve these problems, utilizing the management technology based on Case-Based Reasoning(CBR). CBR refers to the process of solving a new problem on the basis of the resolution of similar past problems. The process which uses the information about past cases and makes a decision for the time being of incident is that you are using the CBR. To do so, explore whether the current problem is consistent with the practice and experience that have been seen in the past. And then if there is a match, based on the past state, it presents a solution. All reasoning is that it has experienced in the past or is made based on past cases that have been accepted by hard training. The degree of equipment aging is steadily accumulated from past times rather than abruptly occurs. If it were outside the normal condition or occurs specific patterns, then the failure would occur. In this paper, in order to consider the case in performing the analysis, it assumes three states of the equipment. And this paper intended to extract a case suitable for the health diagnosis and prediction of equipment through a statistical validation.

2.2. Random Forest

Random Forest(RF) is developed by (Beriman, 2001). Rather than one technique of traditional decision-tree, it is a tree-based ensemble technique was extended to a plurality of wood. Configure the N different training data sets with one another via restoring extracted from the same training data set to form a decision-tree that constitutes the RF. The accuracy of N individual decision-tree may be reduced. However, when the result is summarized from the N number of decision-tree through the randomization techniques such as the bagging or Randomized node optimization, safety and accuracy become high. Indirectly, by using the RF it is possible to measure the importance of each variable in classification and regression problems. In the process of the RF, out-of-bag(OOB) error is determined about each of data point.OOB data refers to data that do not belong to any of the training data set at the time of duplicate extracted with bootstrapping. OBB-Error is defined as error between the actual value and the predicted results that came out from the decision-tree. The maximum advantage of RF has is capable of handling thousands of input variables without variable selection. In addition, it shows the higher classification performance in multiple classes and class-imbalanced data

III. EXPERIMENTS

3.1. Data Description

It was analyzed using the data collected during a period of about 3-4 months from the vacuum pump sensor. Observed data by the sensor is composed of seven variables. Based on the Breakdown Maintenance(BM) point, we make a dataset between the BM point as Run-to-Fault cycle. A total of 35 datasets are created. **Fig.1** shows the dataset of one of the 35 datasets. Show different patterns according to the characteristics of the sensor. Subtle changes in the patterns can be observed in the green and red regions which is after BM and before BM. Initial state of the green region is the state after the BM is occurred. It is

defined as the highest health index and re-configure the system.

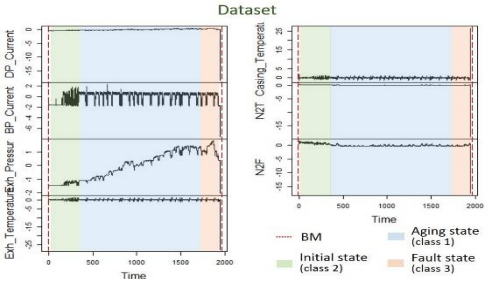


Fig.1. Dataset Plot

Aging state of the blue area is a section showing the increase of aging while the equipment is being operated. It is defined to be continued until fault occurs.Finally, Fault state of red region is in a section before BM occurs. It is defined that this state appears abnormal pattern and sensor values over threshold.In order to perform the classification analysis, this paperdefines the three state as class e.g. Initial state is class1, Fault state is class2, Aging state is class3.The machine learning techniques are performed using the Random Forest which are known to be found good performance for multiple classes and class-imbalanced data(Brown et al, 2012).

3.2. Design of experiments

In this paper, in order to analyze the two experimental conditions, select the 34 datasets among the 35 dataset to create a train&validation dataset and then create a test dataset for the remaining one of the datasets. This paper conducted an experiment to find the ratio of the optimal length for states in the 2-Class and 3-Class respectively using train & validation dataset. It set the various combination of each length ratio for the experiment such as **Table 1** and **Table 2**. To improve the reliability of each time the experiment was repeated 5 times with 5-fold cross validation, and derived the optimal length ratio. After assuming test data set to actual dataset, it is applied to the 2-Class and 3-Class. Then we selected a model appeared superior performance. It was repeated 35 times with Leave one out cross validation (LOOCV) in order to improve the reliability. Finally, it calculated the probability of belonging to Fault state over time and were used as the healthindicator. We used the classification of Accuracy and M.AUC(the Area Under a ROC Curve in multiple class) as a performance measure of this model. It is a multiple class that three states have been considered rather than the binary class. After calculating the AUC for each class, We calculatedM.AUC from the average AUC which comes out in each class.

3.3. Result of Experiments

In **Table1** and **Table 2**,theseshow the experimental results of the 2-Class and 3-Class.Performance measure in each table is the result of 5-Fold cross validation that 5 repetitions were performed for each combination of the state length ratio. Accuracy and M.AUC represents the mean of 5 experiments. **Fig.2** is expressed in a graph the results of the Accuracy and the M.AUC of the **Table 1** and **Table 2**. Looking at the results of the graph, when 2-Class is (Initial + Aging): Fault = 0.8:0.2 and 3-Class is Initial: Aging: Fault = 0.05: 0.9: 0.05 show a high degree of performance. So it is possible to select the optimal length ratio. Overall, if the section of Fault state is close to the BM point, the Accuracy increase because the probability of misclassifying the Fault state into Aging state is reduced. However, in order to avoid the problem of over fitting problem and determine the optimal ratio of the state interval, we consider the M.AUC together. Based on the state length ratio from the above results, we calculate the performance measure of a 3-Class and 2-Class using test dataset. Using LOOCV the Accuracy and M.AUC were calculated and evaluated for 35 test dataset. The results of the experiment considering the Accuracy and M.AUC shows a higher precision and a small standard deviation in 3-Class.Therefore when dataset is divided into three states (Initial, Aging, Fault) and the ratio of the section to be set to Initial: Aging: Fault = 0.05:0.9:0.05, it can be said to be effective in diagnosis and prediction of the equipment failure. **Fig.3**, **Fig.4**, **Fig.5** show the probability that each point is belong to the Fault state for test datasets.

The red line shows the fit line to the predicted probability using the loess regression. In **Fig.3**and**Fig.4** results indicate a pattern that probability increases closer to the BM point. In **Fig.5** the probability in the middle significantly increased. This can know that there was the problem similar to Fault state in the middle of the process. From the results of **Fig.3**, **Fig.4**, **Fig.5**, it is possible that it is able to see the probability of belonging to a fault state as the health indicator.

Table1: number of cases of 2-Class

Num	Initial+Aging	Fault	Accuracy	M.AUC
1	0.60	0.40	0.933	0.933
2	0.70	0.30	0.950	0.947
3	*0.80*	*0.20*	*0.962*	*0.952*
4	0.85	0.15	0.962	0.939
5	0.90	0.10	0.968	0.935
6	0.95	0.05	0.974	0.916
7	0.99	0.01	0.992	0.904

Table2: number of cases of 3-Class

Num	Initial	Aging	Fault	Accuracy	M.AUC
1	0.1	0.8	0.1	0.925	0.940
2	0.1	0.7	0.2	0.922	0.951
3	0.1	0.6	0.3	0.917	0.946
4	0.1	0.5	0.4	0.904	0.940
5	*0.05*	*0.9*	*0.05*	*0.947*	*0.944*
6	0.05	0.8	0.15	0.938	0.948
7	0.05	0.7	0.25	0.933	0.950
8	0.05	0.6	0.35	0.919	0.949
9	0.05	0.5	0.45	0.912	0.936
10	0.1	0.8	0.1	0.929	0.947
11	0.1	0.7	0.2	0.925	0.952
12	0.1	0.6	0.3	0.912	0.945
13	0.1	0.5	0.4	0.904	0.940
14	0.2	0.7	0.1	0.911	0.946
15	0.2	0.6	0.2	0.910	0.951
16	0.2	0.5	0.3	0.901	0.944
17	0.3	0.6	0.1	0.907	0.949
18	0.3	0.5	0.2	0.899	0.951

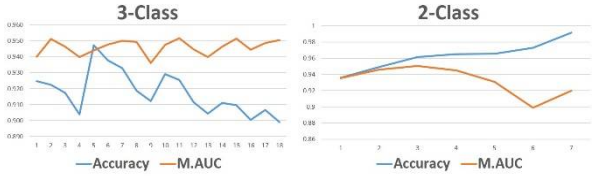


Fig.2. Accuracy and M.AUC Plot

Table 3: Compare 2-Class with 3-Class

Performance Measure	2-Class		3-Class	
	Mean	SD	Mean	SD
Accuracy	0.74	0.20	0.81	0.15
M.AUC	0.67	0.20	0.73	0.16

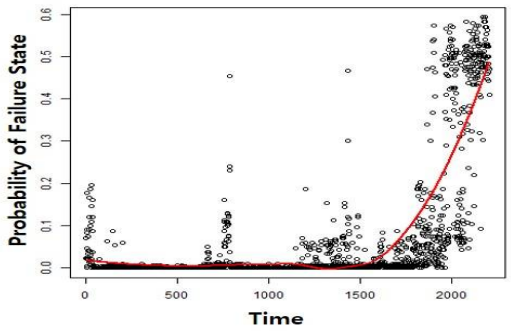


Fig.3. Probability of Fault State(dataset 23)

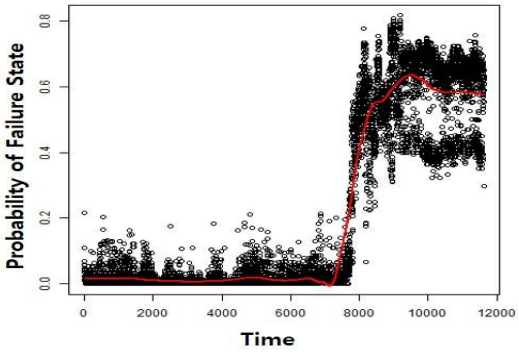


Fig.4. Probability of Fault State(dataset 28)

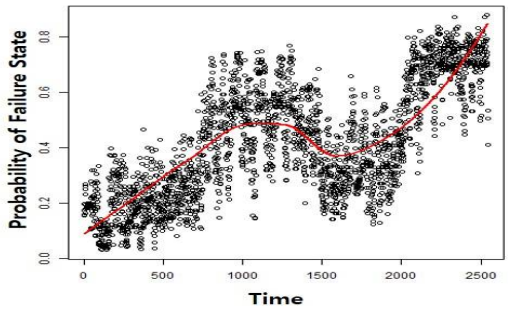


Fig.5. Probability of Fault State(dataset 13)

CONCLUSIONS

In this paper, we propose a classification model that is able to reflect the current state of the equipment through an analysis of the equipment history data collected in real-time in the semiconductor field. The proposed method sets various experimental conditions to define the state of the facility and find the optimal state. And then the analysis is carried out by using the method of machine learning. As the final result, a health index is formed by calculating the probability about fault state per time. This method is expected to be useful in the field of abnormal diagnosis and prediction of facilities in the future.
However, this research needs further research plan to enhance prediction power of health index. For this, extracting additional feature from existing data is needed. It requires additional features extracted from the existing data. Also there is a need to improve the research results to the failure probability can be successfully expressed through comparison of various machine learning techniques besides RF

ACKNOWLEDGMENTS

This work was supported by the BK21 Plus (Big Data in Manufacturing and Logistics Systems, Korea University).
This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2013R1A1A2010019).

REFERENCES

.

[1] Mosallam, Ahmed, Kamal Medjaher, and Noureddine Zerhouni. "Nonparametric time series modelling for industrial prognostics and health management." The International Journal of Advanced Manufacturing Technology 69.5-8 (2013): 1685-1699.
[2] Coble, Jamie, and J. Wesley Hines. "Identifying optimal prognostic parameters from data: a genetic algorithms approach." Annual Conference of the Prognostics and Health Management Society, San Diego, CA, September. 2009.
[3] Coble, Jamie, and J. Wesley Hines. "Applying the general path model to estimation of remaining useful life." International Journal of Prognostics and Health Management Volume 2 (color) (2011): 71.
[4] Barbieri, Federico, et al. "Sensor-Based Degradation Prediction and Prognostics for Remaining Useful Life Estimation: Validation on Experimental Data of Electric Motors." International Journal of Prognostics and Health Management (2015)
[5] Ali, Jaouher Ben, et al. "Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals." Applied Acoustics 89 (2015): 16-27.
[6] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
[7] Brown, Iain, and Christophe Mues. "An experimental comparison of classification algorithms for imbalanced credit scoring data sets." Expert Systems with Applications 39.3 (2012): 3446-3453.

★★★