Performance Analysis of Complex Manufacturing Process with Sequence Data Mining Technique

Kittisak Kerdprasop and Nittaya Kerdprasop

Data Engineering Research Unit,
School of Computer Engineering,
Suranaree University of Technology, Thailand

{kerdpras, nittaya}@sut.ac.th

Abstract

In this paper, we present a sequence analysis method, which is one of the advanced data mining techniques, to identify and extract unique patterns from wafer manufacturing data. Wafer fabrication in the semiconductor industry is one of the most complex manufacturing processes. For such highly complicated operations, maintaining high yields through the statistical process control as a sole monitoring method for quality control is obviously inefficient. We thus investigate the intelligent and semi-automatic technique to help industrial engineers analyzing their production data. Our proposed method has the ability to induce patterns that can reveal and differentiate low performance processes from the normal ones. We also provide program coding of the proposed sequence analysis method, implemented with the R language, for easy experimental repetition.

Keywords: *Sequence analysis, Performance pattern, Intelligent manufacturing, Sequence data mining, R programming language*

1. Introduction

Sequence is an ordered set of elements in which each element can be numerical, categorical, or a mixture of attributes. The order of elements could be determined by their occurring time or positions. If the order is by time and the elements of a sequence are real values, it is a time series. When the sequence elements are discrete, it is a categorical sequence [13]. Sequence mining is a recently active field of research in knowledge discovery and data mining. The applications of the available techniques are mostly in the areas of bioinformatics and financial analysis. In this paper, we demonstrate the potential application of sequence data mining to discover the operational sequences of tools causing low performances in the semiconductor manufacturing process.

Semiconductor manufacturing is a highly complex production process composed of hundreds of steps. The major processes in most semiconductor industries are: production of silicon wafers from pure silicon material, fabrication of integrated circuits onto the raw silicon wafers, assembly by putting the integrated circuit inside a package to form a ready-to-use product, and testing of the finished products [9]. A constant advancement in the semiconductor industry is due mainly to persistent improvement of the wafer fabrication process.

The fabrication process consists of a series of steps to cover special material layers over the wafer surface. Wafers re-enter the same processing machines as each layer is successively covered. Some defects in this complicated process can make the final products fail the test. Early fault detection during this critical manufacturing process can obviously improve product quality and reliability.

Recent trend in intelligent manufacturing is to apply the data mining techniques to automatically identify patterns and causal relationships leading to poor yield [2, 16]. In this paper, we expand the frontier of data mining application to the manufacturing process area by proposing an advanced sequence data mining technique. Our proposed technique can be viewed as a compliment of the classical statistical process control in that it can help engineers detecting process variations in a semi-automatic manner. In this paper, we demonstrate the potential application of sequence data mining technique implemented with the R language [18] to discover the operational sequences of tools causing low yields in the complex fabrication process.

2. Related Work

In recent years, many manufacturing tools are equipped with sensors to facilitate real-time monitoring of the production process [3]. These tool-state and production-state sensor data provide an opportunity for efficient control and optimization. Unfortunately, such measurement data are so overwhelming that timely detection of any fault during the production process is difficult. Therefore, automatic and advanced process control method is required.

Ison and colleagues [8] proposed a decision tree classification model to detect fault of plasma etch equipment. The model was built from the five sensor signal data. Goodlin *et al.*, [6] proposed to build a specific control chart for detecting specific type of faults. They collected tool-state data directly from the etcher. These data consist of 19 variables. The work of Spitzlsperger and colleagues [14] was also based on the statistical method. They adopted the multivariate control chart method to maintain changes in the mean and standard deviation coefficients by re-modelling technique.

Later interest in fault detection has been shifted toward the non-parametric approaches. He and Wang [7] proposed to use the k-nearest neighbor rule for fault detection. Verdier and Ferreira [17] also applied the k-nearest neighbor method, but they proposed to use the adaptive Mahalanobis distance instead of the Euclidean distance. Tafazzoli and Saif [15] proposed a combined support vector machine methodology for process fault diagnosis. Ge and Song [5] applied support vector data to the principal component analysis method to detect process abnormalities.

Most work on fault detection methods has studied the process control problem with a few features of tool-state and process-state measurement data. McCann and his team [10] proposed a rather different setting in which the measurement data from the wafer fabrication process contain as much as 590 features. They applied feature selection technique to select only 40 features for further analysis.

In this work, we apply a data mining technique that can handle 300 features of sequential data, rather than independent and discrete data as proposed in all the previous work. Sequence data mining of manufacturing process appeared in the literature just a few years ago [11, 12]. Our work presented in this paper is different from others in that we apply sequence analysis as an exploratory tool, instead of the classification tool. Moreover, we adopt the open source paradigm for the purpose of re-experimentation.

3. Sequence Analysis Method

3.1. Manufacturing Process Data

In our sequence analysis, we use the dataset named SETFI (SEmiconductor Tool level Fault Isolation), which is a simulated dataset [1] that closely mimics the actual high complexity of semiconductor manufacturing process. The dataset contains 4000 records of the wafer fabrication process. During the process each, a wafer goes through sequence of operations in batch, which is called lot in this dataset. The sequences of hundreds of operations might be different from lot to lot, but these operations involve only twenty tools, number 1 to 20. At each operation unit, only a single tool is in operation. Tool distribution in the wafer fabrication process is graphically shown in Figure 1.

At the end of the fabrication process, a number of inspection steps are carried out to measure the product performance. Wafer lots that fail the inspection tests need re-processing. Low performance metric is often caused by a small subset of tools. Identifying such problematic tools at an early stage can obviously improve yield performance of the semiconductor manufacturing. Some data instances of the SETFI dataset are shown in Table 1.

The original SETFI dataset contains the tool number applied in each of the 300 operational units together with the timestamps of each operation. In this sequence analysis study, we remove the first column (Lot#) because it plays no role to the discovering of sequence patterns. We also ignore the timestamps because our main objective is the categorical sequence analysis, not a time series analysis.

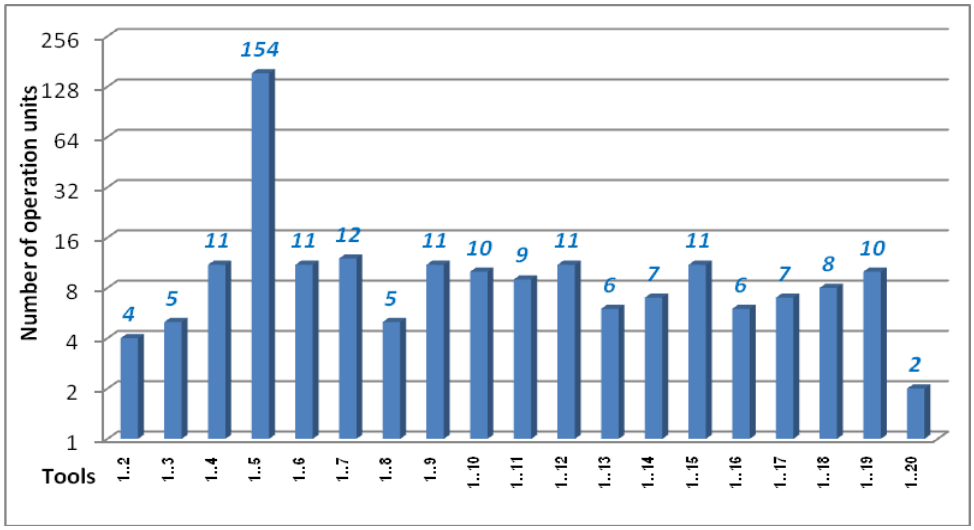

Figure 1. Distribution of Tools Applied in the Wafer Operational Units

Table 1. Example of Data Instances in the SETFI Dataset

Lot#	Op_1	...	Op_{300}	T_1	...	T_{300}	Performance
3699	2	…	3	77.69978	…		2841.763
1427	9	…		31.51063	…	320.6053	2779.744
…	…	…	…	…	…	…	…
1753	7	…	5	36.4376	…	328.7119	2732.957

The meaning of each data instance in Table 1 can be explained as follows. The first data instance contains information of a wafer lot number 3699 that starts the fabrication process with a tool number 2 and ends with a tool number 3. Its performance metric is 2841.763. The second data instance shows record of a wafer lot number 1427. It firstly goes through the tool number 9, but the tool number of its last operational unit (Op_{300}) is missing. Missing values in this dataset are around 25%.

From the manufacturing process dataset that contains information of tools applied in the 300 operational units of 4000 wafer lots, we designed the performance sequence analysis framework as illustrated in Figure 2.

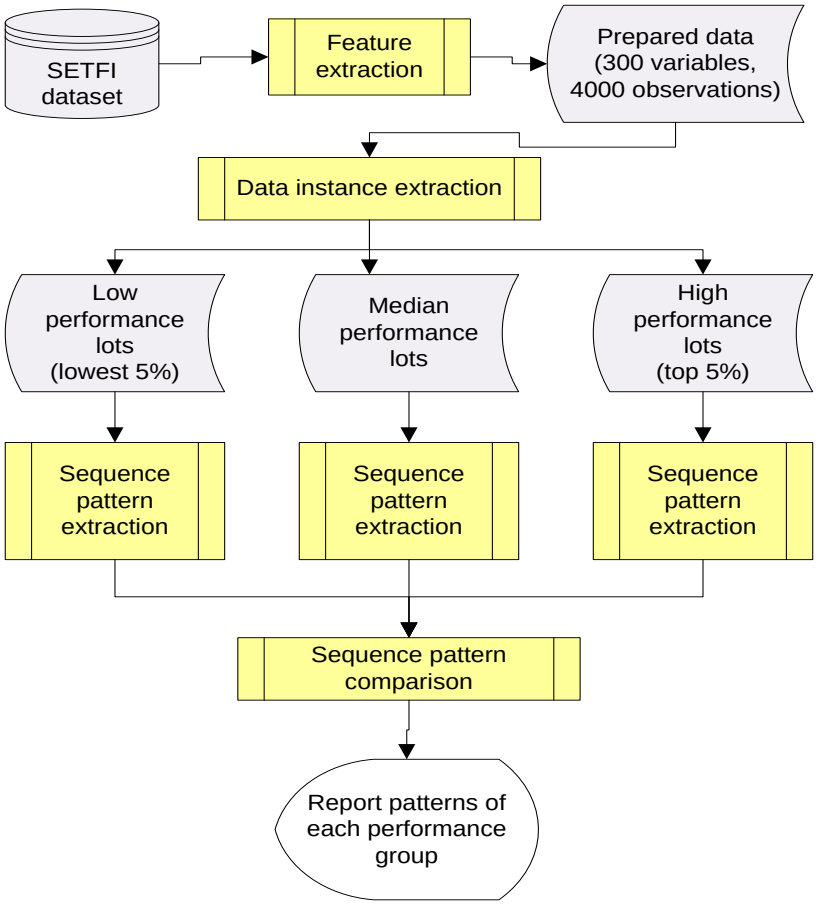

Figure 2. A Methodology for the Manufacturing Sequence Analysis

Table 2. Performance of Wafer Subgroups

Wafer lot subgroup	Performance			
	Maximum	Minimum	Average	S.D.
Low (200 lots)	2574.012	2177.438	2503.816	66.95
Median (201 lots)	2790.671	2778.334	2784.345	3.70
High (201 lots)	3293.183	2992.259	3062.469	63.95
All (4000 lots)	3293.183	2177.438	2787.924	125.84

The first step of data preparation for our analysis method is to extract features (or variables) containing the tools used in the 300 operational units together with the performance metric, which is the last column in the SETFI dataset.

We then divided the dataset into three subgroups: low, median, and high performance lots. Each subgroup contains approximately 200 to 201 data instances. Performance statistics (maximum, minimum, average values and standard deviation within the subgroup) of the three subgroups are summarized in Table 2.

3.2. Performance Pattern Mining Technique

Data division into three subgroups and sequence extraction are performed by a computer program, implemented with the R language. Program coding is provided in Figure 3.

The program calls seqdef, seqecreate, and seqefsub functions from the library TraMineR [4]. There is only one function in the program, which is mainT. The SETFI dataset is in the comma-separated-value (csv) format and stored in the folder sequence-mining. The first command in the program is to read the data and store in the variable 'dat'. The first column, which is the lot number, is then removed. Then the dataset has been sorted in descending order according to the performance value. The ordered data of 4000 wafer lots are called 'dat2all'. This dataset is divided into three subsets: 'data2low', 'data2mid', and 'data2top'.

If the function mainT() has been called with no parameter, the 'data2low' will be processed by default to search for sequences of a low performance subgroup. To extract sequence patterns from other subsets, the parameter has to be specified. For example, mainT(high) is a command to extract patterns from a group of wafer lots with high performance, and mainT(mid) is for median subgroup analysis.

The parameters 'from' and 'to' are for identifying range of data columns to be analyzed. Parameter 'per' is a percentage to split data into high, low, and median subgroups. The last parameter is minimum support value, 'min', in which the value 0.3 has been set as default. Users can change these parameters upon the function call. To analyze wafer lot patterns, we have to run this program three times, *i.e.*, each execution for each data subset. Sequences of all 4000 wafer lots are also induced for comparative analysis.

```
library(TraMineR)
mainT <- function(from=1, to=300,lot='low', per=0.05, min=0.3)
{  dat  <- read.csv('C:/sequence-mining/SETFI.csv')
   dat <- dat[-1]       # remove first column
   dat2 <- dat          # make a copy of dataset
   dat2all <- dat2[with(dat2,order(res)),]
   dat2low <- dat2all[1:(per*nrow(dat2)),]
   dat2mid <- dat2all[(0.5*nrow(dat2)-(per*nrow(dat2))/2) : (0.5*nrow(dat2)+(per*nrow(dat2))/2),]
   dat2top <- dat2all[((1-per)*nrow(dat2)):4000,]
   if (lot=='low') dat2 <- dat2low
   if (lot=='high') dat2 <-dat2top
   if (lot=='all') dat2 <-dat2all
   if (lot=='mid') dat2 <-dat2mid
   mvad.seq <- seqdef(dat2, var=from:to, missing=NA)
  #  Event sequence analysis
   mvad.seqe <- seqecreate(mvad.seq)
   fsubseq <- seqefsub(mvad.seqe, pMinSupport = min )
   print(fsubseq[1:50])
  # plot the 15 most frequent sequences
  plot(fsubseq[1:15],  main=paste(nrow(dat3),'records at',lot,'lot ,Columns :Col', from,'-',to))
}
```

Figure 3. A Program for Manufacturing Sequence Analysis

4. Sequence Analysis Results

To analyze the wafer fabrication lot patterns, we have to run the sequence extraction program four times, varying a lot parameter as ‘low’, ‘mid’, ‘high’, and ‘all’ in each execution. Users may call these executions in one time and save the output in a file ‘out.txt’ with the following commands:

```
sink('out.txt')
for(t in c('low', 'mid', 'high', 'all')) { mainT(lot=t) }
sink()
```

The results of all executions are to be written in a file ‘out.txt’. The interpretation of discovered subsequence extraction results is straightforward. For instance, the subsequence displayed as (2>5) means the tool number 2 had been applied prior to the tool number 5.

The focus of our sequence analysis is the ability to differentiate frequent manufacturing process patterns of low performance lots from the high performance lots.

From the clustering results, we concentrate our analyses on the bottom low lots (containing 74 records) that show the lowest performance values comparative to the top high lots (with 46 records) that show the highest performance metrics. The sequences of these two groups are shown in Figure 4.

Long chaining subsequences such as (1>3)-(3>2)-(2>1) can be interpreted as the sequence of tool number 1 followed by tool number 3 used in the two operational units is normally preceding the other two tool sequences, that is, 3→2 and 2→1, respectively. In our experimentation, we set the number of sequence items to be at most three (due to the memory space limitation), and to display only the top-40 sequences.

	(Low) Subsequence		*(High)* Subsequence
1	(1>4)	1	(1>3)
2	(1>4)-(4>3)	2	(1>3)-(2>1)
3	(1>5)	3	(1>3)-(2>3)
4	(1>5)-(2>4)	4	(1>3)-(3>2)
5	(2>3)	5	(1>3)-(3>2)-(2>1)
6	(2>3)-(3>1)	6	(1>3)-(3>2)-(2>3)
7	(2>3)-(3>2)	7	(1>3)-(3>2)-(3>2)
8	(2>3)-(4>3)	8	(1>3)-(3>2)-(5>3)
9	(2>4)	9	(1>3)-(3>4)
10	(2>4)-(2>4)	10	(1>3)-(3>4)-(2>1)
11	(2>4)-(2>5)	11	(1>3)-(3>4)-(3>2)
12	(2>4)-(2>5)-(2>4)	12	(1>3)-(4>1)
13	(2>4)-(2>5)-(5>2)	13	(1>3)-(5>2)
14	(2>4)-(3>1)	14	(1>3)-(5>2)-(2>1)
15	(2>4)-(3>2)	15	(1>3)-(5>3)
16	(2>4)-(3>4)	16	(1>4)
17	(2>4)-(5>2)	17	(1>4)-(1>3)
18	(2>4)-(5>3)	18	(1>4)-(1>3)-(3>2)
19	(2>5)	19	(1>4)-(2>1)
20	(2>5)-(1>5)	20	(1>4)-(2>1)-(3>2)
21	(2>5)-(2>3)	21	(1>4)-(2>3)
22	(2>5)-(2>4)	22	(1>4)-(2>3)-(3>2)
23	(2>5)-(2>4)-(3>2)	23	(1>4)-(2>4)
24	(2>5)-(3>2)	24	(1>4)-(2>5)
25	(2>5)-(5>2)	25	(1>4)-(2>5)-(2>3)
26	(3>1)	26	(1>4)-(2>5)-(3>2)
27	(3>1)-(2>3)	27	(1>4)-(2>5)-(4>5)
28	(3>1)-(4>2)	28	(1>4)-(3>2)
29	(3>1)-(5>2)	29	(1>4)-(3>2)-(4>2)
30	(3>2)	30	(1>4)-(3>5)
31	(3>2)-(1>4)	31	(1>4)-(4>1)
32	(3>2)-(2>3)	32	(1>4)-(4>1)-(2>3)
33	(3>2)-(2>3)-(3>1)	33	(1>4)-(4>2)
34	(3>2)-(2>3)-(3>2)	34	(1>4)-(4>2)-(2>1)
35	(3>2)-(2>4)	35	(1>4)-(4>2)-(2>4)
36	(3>2)-(2>4)-(3>2)	36	(1>4)-(4>2)-(4>5)
37	(3>2)-(3>1)	37	(1>4)-(4>5)
38	(3>2)-(3>1)-(5>2)	38	(1>4)-(4>5)-(3>2)
39	(3>2)-(3>2)	39	(1>4)-(4>5)-(4>2)
40	(3>2)-(3>2)-(5>3)	40	(1>4)-(5>3)

Figure 4. Frequently Occurred Tool Sequences of Low versus High Performance Lots

To compare the highly occurred tool sequences of low performance lots against the high performance lots, we graphically draw the diagrams (Figure 5) of top-14 tool sequences. From the diagrams, we can notice that the top performance lots involve sequences of tool numbers 1, 2, and 3, whereas the low performance sequences involve the tool numbers 1 and 5.

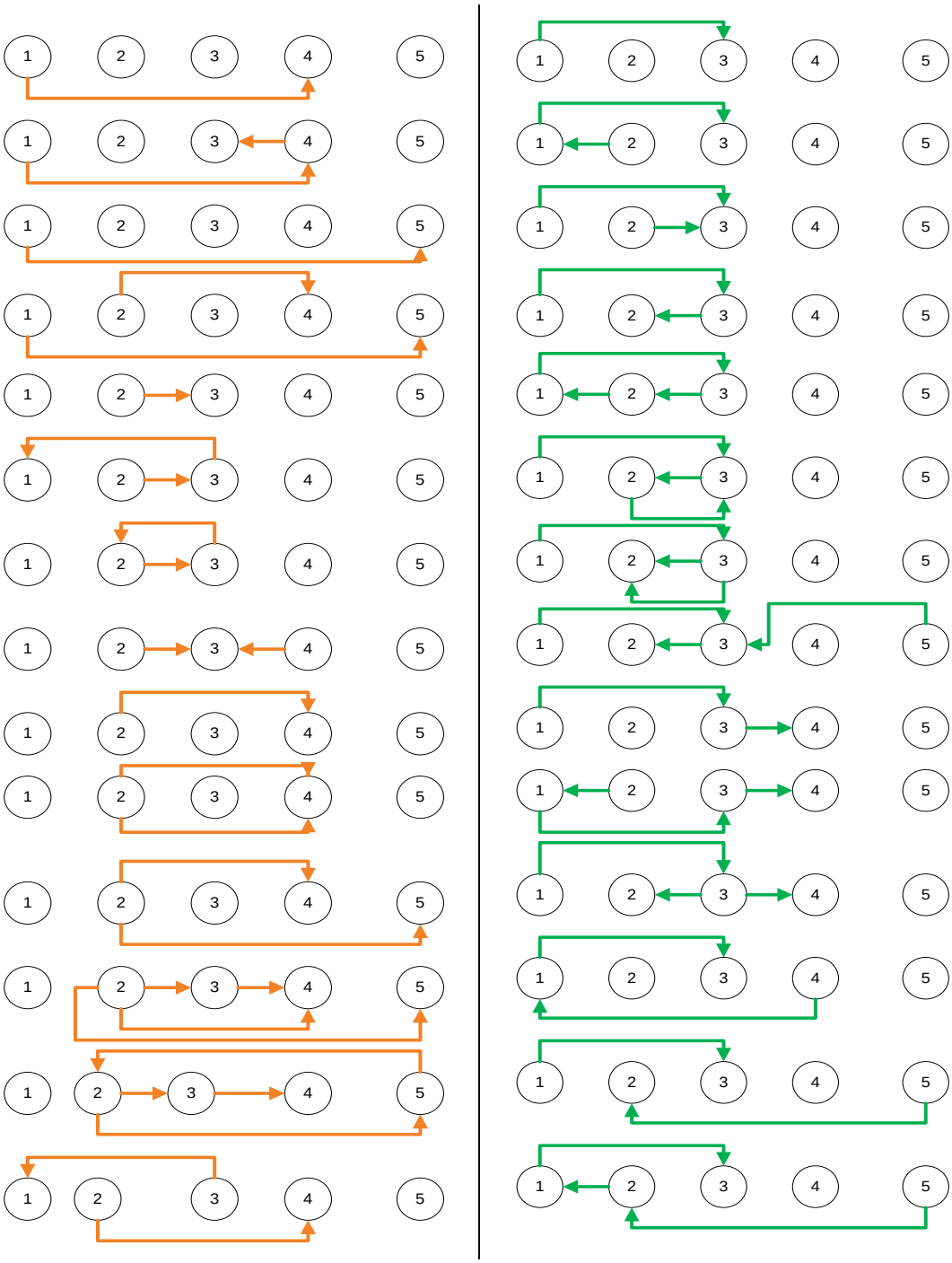

Figure 5. Top-14 Sequences of Low (left) against High (right) Performance Wafer Lots

We then decompose the chaining sequences of the top-50 sequences in the three subgroups (that are low, median, and high performance wafer lots) down to a single sequence to find a unique sequence in the low performance and the high performance group. The outcome is shown in Figure 6. We can draw a conclusion from this experiment that a unique pattern in the low performance group is a sequence of tool 5→1, and a unique pattern in the high performance group is 1→3.

Low Performance	Median Performance	High Performance
		1→3
1→4		1→4
1→5	1→5	
	2→1	2→1
2→3	2→3	2→3
2→4		2→4
2→5	2→5	2→5
3→1	3→1	
3→2	3→2	3→2
3→4	3→4	3→4
	3→5	3→5
	4→1	4→1
4→2	4→2	4→2
4→3	4→5	4→5
5→1		
5→2	5→2	5→2
5→3	5→3	5→3

Figure 6. A Comparison of Unique Patterns among Subgroups of Wafer Lots

5. Conclusion

Most highly complex manufacturing industries have produced constantly hundreds of metrology data that are awaiting for process engineers to analyze for the purpose of maintaining efficient operations and getting optimum yield of high quality products. For such a large volume of measurement data, automatic data analysis technique is essential. We thus propose the application of sequence data mining technique to help engineers analyzing problematic process sequences in the semiconductor industries.

We designed an analysis framework to group operational process data into three categories: processes with low, high, and moderate performance metrics. Process data in each category were then analyzed with the sequence mining program written in R language. We found from the experimental results that the frequently occurred sub-sequences of each category show some unique patterns. Sequence analysis technique presented in this paper is semi-automatic in the sense that unique pattern inspection has to be done by human. We thus plan to further our research towards the design and implementation of an automatic tool to timely detect process trends leading to low performance products.

Acknowledgements

This research was supported by the SUT Research and Development Fund, Suranaree University of Technology.

References

[1] AA&YA, Intel, SETFI: Manufacturing data: Semiconductor tool fault isolation, Causality Workbench Repository, [Online] Available: http://www.causality.inf.ethz.ch/repository.php **(2008)**
[2] T. Chen, "An evolving fuzzy-neural rule with slack diversification for enhancing job dispatching", International Journal of Grid and Distributed Computing, vol. 5, no. 2, **(2012)**, pp. 1-8.
[3] T. Chen and Y. Wang, "Estimating the job cycle time in wafer fabrication with distributed sensors", International Journal of Hybrid Information Technology, vol. 5, no. 12, **(2012)**, pp. 81-88.
[4] A. Gabadinho, G. Ritschard, M. Studer and N. Muller, "Extracting and rendering representative sequences", in A. Fred, J.L.G. Dietz, K. Liu, and J. Filipe (eds.), Knowledge Discovery, Knowledge Engineering and Knowledge Management. Series: Communications in Computer and Information Science (CCIS), Springer, vol. 128, **(2011)**, pp. 94-106.
[5] Z. Ge and H. Song, "Semiconductor manufacturing process monitoring based on adaptive substatistical PCA", IEEE Transactions on Semiconductor Manufacturing, vol. 23, no. 1, **(2010)**, pp. 99-108.
[6] B. E. Goodlin, D. S. Boning, H. H. Sawin, H. H. and B. M. Wise, "Simultaneous fault detection and classification for semiconductor manufacturing tools", Journal of The Electrochemical Society, vol. 150, no. 12, **(2003)**, pp. G778-G784.
[7] Q. P. He and J. Wang, "Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes", IEEE Transactions on Semiconductor Manufacturing, vol. 20, no. 4, **(2007)**, pp. 345-354.
[8] A. M. Ison, W. Li and C. J. Spanos, "Fault diagnosis of plasma etch equipment", Proceedings of IEEE International Symposium on Semiconductor Manufacturing, San Francisco, **(1997)**, pp. B-49-B-52.
[9] G. S. May and C. J. Spanos, "Fundamentals of Semiconductor Manufacturing and Process Control", John Wiley & Sons, **(2006)**.
[10] M. McCann, Y. Li, L. Maguire and A. Johnston, "Causality challenge: benchmarking relevant signal components for effective monitoring and process control" Proceedings of JMLR Workshop, Canada, **(2008)**, pp. 277-288.
[11] L. Rokach, R. Romano and O. Maimon, "Mining manufacturing databases to discover the effect of operation sequence on the product quality", Journal of Intelligent manufacturing, vol. 19, **(2008)**, pp. 313-325.
[12] N. Ruschin-Rimini, O. Maimon and R. Romano, "Visual analysis of quality-related manufacturing data using fractal geometry", Journal of Intelligent manufacturing, vol. 23, **(2012)**, pp. 481-495.
[13] S. Sarawagi, "Sequence data mining", S. Bandyopadhyay, U. Maulik, L. B. Holder and D. J. Cook (eds.), Advanced Methods for Knowledge Discovery from Complex Data, Springer, **(2005)**, pp. 153-187.
[14] G. Spitzlsperger, C. Schmidt, G. Ernst, H. Strasser and M. Speil, "Fault detection for a via etch process using adaptive multivariate methods", IEEE Transactions on Semiconductor Manufacturing, vol. 18, no. 4, **(2005)**, pp. 528-533.
[15] E. Tafazzoli and M. Saif, "Application of combined support vector machines in process fault diagnosis", Proceedings of American Control Conference, St. Louis, **(2009)**, pp. 3429-3433.
[16] C. Teutsch, D. Bernt, J. Schnee, M. Hubner and N. Bachfischer, "Optical inspection of laser markings in the production process", International Journal of Database Theory and Application, vol. 3, no. 3, **(2010)**, pp. 39-47.
[17] G. Verdier and A. Ferreira, "Adaptive Mahalanobis distance and k-nearest neighbor rule for fault detection in semiconductor manufacturing", IEEE Transactions on Semiconductor Manufacturing, vol. 24, no. 1, **(2011)**, pp. 59-68.
[18] Y. Zhao, R and Data Mining: Examples and Case Studies, [Online] Available: http://www.rdatamining.com **(2012)**.

Authors

Kittisak Kerdprasop is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer Science from Nova Southeastern University, U.S.A., in 1999. His current research includes Data mining, Artificial Intelligence, Functional and Logic Programming Languages, Computational Statistics.

Nittaya Kerdprasop is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in Radiation Techniques from Mahidol University, Thailand, in 1985, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer Science from Nova Southeastern University, U.S.A, in 1999. Her research of interest includes Knowledge Discovery in Databases, Artificial Intelligence, Logic Programming, and Intelligent Databases.