Deep Learning Based Virtual Metrology and Yield Prediction in Semiconductor Manufacturing Processes

Myong Kee Jeong[1], Jeongsub Choi[1], Youngdoo Son[2], and Jihoon Kang[3]

[1]*Industrial and Systems Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, 08854, USA*

mjeong@rutgers.edu
jeongsub.c@rutgers.edu

[2] *Industrial and Systems Engineering, Dongguk University, Seoul, 04620, South Korea*

youngdoo@dongguk.edu

[3]*Algorithm Research team, Samsung SDS, Seoul, 06765, South Korea*

jh.82.kang@samsung.com

ABSTRACT

We present a deep learning based supervised autoencoder to extract meaningful features from massive in-line sensor functional signals of semiconductor manufacturing processes. Based on those extract features, we build the virtual metrology model to predict important quality characteristics of the process and the yield prediction model. A real-life case is studied in this work, and the empirical results show that the proposed model outperforms general approaches for the predictions using signal data.

1. INTRODUCTION

Virtual metrology (VM) is a useful tool to define the relationships between process data and metrology data. Reliable measures from VM techniques for physical metrologies lead to cost reduction caused by prospective measurements, and production quality improvements. In semiconductor manufacturing, VM techniques are employed not only for real-valued prediction of physical metrologies, but also for fault detection and classification.

In recent years, deep learning models have been attracted to researchers wherein the models show successful performance on machine learning tasks such as speech recognition, image classification, and natural language processing. Also, many studies with deep learning models have been conducted in manufacturing applications (Lee et al. 2017). However, most of the studies focus on classification tasks.

We propose a new VM method based on a deep learning model. The proposed model aims to extract underlying features to predict real-value quality characteristics and a yield rate using process signal data from multiple sensors. Specifically, a goal is to predict the target values using raw signals without elaborate signal data preprocessing such as signal alignment. In the proposed model, the regularization on individual weights is employed for the model sparsity, and the regularization on the groups of weights is considered for the sparsity of sensors.

2. SUPERVISED AUTOENCODER WITH REGULARIZATION

An autoencoder is an unsupervised learning algorithm that extracts features in a neural network framework reconstructing input values by setting the target values in the model equal to the input values. In the stacked autoencoder (Vincent et al. 2010) with K layers, the d_k-dimensional vector of activations on the k-th layer, $\mathbf{y}^{(k)} \in \mathbb{R}^{d_k}$, for $k = 1, ..., K$ is computed as

$$\mathbf{y}^{(k)} = f_{(k)}\left(\mathbf{z}^{(k)}\right) \quad (1)$$

$$\mathbf{z}^{(k)} = \mathbf{W}^{(k)}\mathbf{y}^{(k-1)} + \mathbf{b}^{(k)} \quad (2)$$

where $f_{(k)}$ is the element-wise activation function for the k-th layer, $\mathbf{W}^{(k)} \in \mathbb{R}^{d_{(k-1)} \times d_k}$ and $\mathbf{b}^{(k)} \in \mathbb{R}^{d_k}$ are the weight matrix, bias vector and respectively, for the k-th layer, and $\mathbf{y}^{(0)}$ indicates an input vector $\mathbf{x}$. The model is formulated as minimizing the follow cost:

$$\mathcal{L}(\theta) = \frac{1}{2}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \quad (3)$$

where θ is a set of all the parameters employed in the model, and $\hat{\mathbf{x}}_i$ is the reconstructed input vector of observation i. To train the model, the back-propagation learning rule is widely employed.

The sparse modeling is successfully implemented in deep learning models in the literature. We consider the

regularization not only on the individual weights in $\mathbf{W}^{(k)}$ for $k = 1, \ldots, K$, but also on the groups of the parameters in the weight matrices. That is, given R groups of the input variables, the weights associated with the groups on the first hidden layer, $\mathbf{V}_r^{(1)}$ for $r = 1, \ldots, R$, respectively, are penalized as $\sum_{r=1}^{R} \lVert \mathbf{V}_r^{(1)} \rVert_F$ where $\mathbf{W}^{(1)} = [\mathbf{V}_1^{(1)} | \cdots | \mathbf{V}_R^{(1)}]$.

Furthermore, a supervised approach is considered to extract features for better performance of prediction in target values. The predicted value of each observation with the features extracted from the autoencoder model is compared to the corresponding target value, and the model is penalized according to the difference between the values.

3. EXPERIMENT

3.1. Virtual Metrology and Yield Prediction

In semiconductor manufacturing, products, i.e. wafers, are processed through a number of microfabrication stages, and, among them, a stage, which is regarded as the most influential on the fabrication of wafers, was chosen by engineering experts. For each wafer, the process equipment is monitored by more than 200 sensors, and 14 critical dimensions (CDs) of the wafer quality are gauged before processing at the subsequent stage. After the completion of all the fabrications, the yield rate of the wafer is computed where the yield rate (YR) is the ratio of non-defective chips at predetermined dies on the wafer.

The dataset employed in this study consists of 298 wafer samples that contain the signals from 85 sensors of process equipment, and a predetermined CD (CD1) and YR of the wafers, accordingly. The process equipment was observed for each wafer during 64 sub-operations where the sub-operations were defined based on the change of the control state in Sensor 1. Figure 1 shows a part of three sample signals from four sensors in the dataset. The variance of the sub-operations' processing times resulted in the different total length of the signals of the wafers. For the model implementation, zero values were padded to the end of the signals whose length are shorter than the longest, 654 time stamps, in order to prevent the loss of information.

3.2. Results

For the experiments, the observations in the dataset were divided into three sets. The VM model and model for YR prediction are trained with a training set of 150 wafers where the parameter(s) of each model wafers is optimized using a validation set of 50 wafers, and the prediction performance of the model is computed with a testing set of 98 wafers. For comparison of model performance, we considered the following models: the Lasso regression (Lasso), the principal component regression (PCR), the Lasso regression with the features from principal component analysis (PC-Lasso), the kernel principal component regression (KPCR), the Lasso with the features from the proposed autoencoder of two hidden layers (AE-Lasso), the PCR with the features from the proposed model (AE-PCR), and the Lasso with the features from PCA of the features from the proposed model (AE-PC-Lasso). The prediction models with the features from the proposed autoencoder model showed better accuracy of prediction both in CD1 and YR.

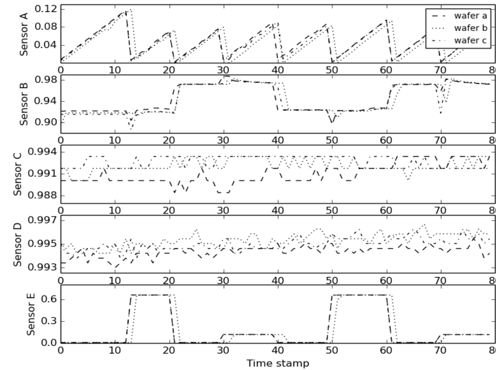

Figure 1. Sample signal data from five process senosors.

Table 1. The results of MSE for VM and YR prediction.

Model	CD1	YR
Lasso	0.0840	0.0020
PCR	0.3056	0.0072
PC-Lasso	0.0846	0.0029
KPCR	0.0780	0.0019
AE-Lasso	0.0369	0.0017
AE-PCR	0.0445	0.0171
AE-PC-Lasso	0.0459	0.0016

4. CONCLUSION

We proposed the supervised autoencoder with regularization for the models of the VM and YR prediction in semiconductor manufacturing using functional signals data of process equipment. The experiments showed that the proposed method resulted in better prediction performance than the other methods.

REFERENCES

Lee, H., Kim, Y., & Kim, C. O. (2017). A Deep Learning Model for Robust Wafer Fault Monitoring with Sensor Measurement Noise. *IEEE Transactions on Semiconductor Manufacturing*, vol. 30(1), pp. 23-31.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, vol. 11, pp. 3371-3408.