An Information Theory-based Approach to Data Clustering for Virtual Metrology and Soft Sensors

GIAN ANTONIO SUSTO
University of Padova
Department of Information Engineering
Via Gradenigo 6/B, 35131 Padova
ITALY
gianantonio.susto@dei.unipd.it

ALESSANDRO BEGHI
University of Padova
Department of Information Engineering
Via Gradenigo 6/B, 35131 Padova
ITALY
beghi@dei.unipd.it

Abstract: Soft Sensors (SSs) are on-line estimators of "hardly to be measured" quantities of a process. The difficulty in measuring can be related to economic or temporal costs that cannot be afforded in a high-intensive manufacturing production. In semiconductor manufacturing this technology goes with the name of Virtual Metrology (VM) systems. While a lot of efforts in research have been produced in the past years to identify the best regression algorithms for these statistical modules, small amount of work has been done to develop algorithms for data clustering of the entire production. This paper contains a new Information Theory-based approach to data clustering for Virtual Metrology and Soft Sensors; the proposed algorithm allows to automatically split the dataset into groups to be equally modeled. The proposed approach has been tested on real industrial dataset.

Key–Words: Semiconductor Manufacturing, Virtual Metrology, Soft Sensor, Data Clustering, Statistical Distance, f-divergence.

1 Introduction

Soft Sensors (SSs) [16] are on-line estimators of "hardly to be measured" quantities of a process. The difficulty in measuring can be related to economic or temporal costs that cannot be afforded in a high-intensive manufacturing production. SSs exploit the availability of process and logistic variables that are always recorded in the process tools, and therefore are "cheap" to be used, for inferring process outputs that are subject to large measurement delays/costs that slow down the production.

SSs technology goes with different names depending on the research community:

- *SS* in chemical industry and biotechnology;
- *Inferential Sensor* in industrial genetics and heating systems;
- *Virtual Sensor* in manufacturing and robotics;
- *Virtual Metrology* (VM) in semiconductor manufacturing [7].

The present work has been performed in the last of the aforementioned areas, semiconductor manufacturing. Modeling of semiconductor manufacturing process, and of several other modern manufacturing, is a challenging task partially due to the huge *data fragmentation*; hundreds/thousands of products are run on the same machine, with different tool settings (called *recipes*); for several processes the dataset is even further complicated by the fact that each product has a different target the equipment is composed of multi-chambers that exhibit different behaviors [11].

Such huge data fragmentation cannot be dealt with by considering separately every specific case, since there are usually not enough data to identify and validate a confident mathematical model for each product. It is therefore necessary to group together data collected under different equipment operating conditions. A smart data clustering can therefore enhance prediction accuracy and it is necessary to model the overall fab production.

While a lot of efforts in research have been dedicated in the past years to identify the best regression algorithms for such statistical modules [10], small amount of work has been done to design algorithms for data clustering of the overall fab production. This paper contains a new Information Theory-based approach to data clustering for VM and, generally, for SSs; the proposed algorithm allows to automatically split the dataset into groups to be equally modeled.

The paper is organized as follows; in Section 2 a brief overview on VM Systems and modeling is provided; in Section 3 the new proposed approach to data clustering is presented, while in Section 4 the proposed approach is tested on an industrial dataset. Finally in Section 5 concluding remarks are provided.

ISBN: 978-1-61804-131-9

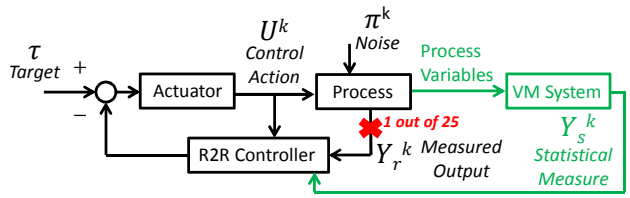

Figure 1: R2R control scheme with Physical and VM (statistical) Measures.

2 Virtual Metrology Systems

As said in the Introduction, collecting measurements in several manufacturing areas for every single process iteration may be hardly feasible due to the corresponding increase of costs and production time. For semiconductor manufacturing processes this is always the case and the current practice is to only monitor the critical dimension (CD) of few wafers in a lot[1].

In the production setup considered in this paper, for instance, only 1 wafer out of 25 is actually measured. Moreover, with few measurements on a lot, equipment-performance drifts between lots are difficult to be promptly detected [3]. A Virtual Metrology (VM) system consists of a mathematical model of the system under consideration for estimating a "costly to measure" physical variable y where tool variables X are used as inputs and it allows to predict values of the relevant variables, without increasing the number of physical measurements by exploiting statistical analysis on tool data and available measurements.

Several semiconductor manufacturing processes benefit of the presence of a Lot-to-Lot (L2L) controller [15]. Based on the physical measurements performed on one wafer in a lot, the process parameters acting on the following lot are updated. The introduction of a VM system may lead to a more accurate, Wafer-to-Wafer (W2W), control policy that allows to detect and reduce the number of faulty wafers as described in Fig. 1. Furthermore, information from VM systems can be exploited to develop Predictive Maintenance modules [14], [12].

It is worth noticing that major European Nanoelectronics Industries are focusing their efforts on developing statistical metrology systems [12, 13] to decrease the number of defective products, increase process stability and even decrease the number of physical measures performed [4]

VM system have been proposed in the literature for Chemical Vapor Deposition (CVD), Etching, and Lithography. Also fab-wide VM structures have been proposed.

[1]For the readers that are not familiar with semiconductor manufacturing procedures, the wafer, a thin slice of silicon used as a substrate for microelectronic devices, is the main product. Production is organized in *lots*, groups of wafers of usually 25 or 50 units.

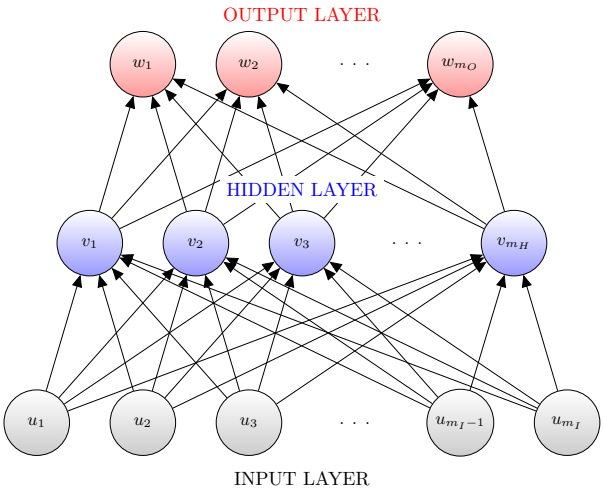

Figure 2: Schematic of a single hidden layer, feed-forward neural network.

2.1 Neural Network-based VM

Given the high complexity of semiconductor processes, black-box approaches to modeling for VM are always preferred to the physical, first-principles ones. Several non-parametric approaches to this problem have been proposed in the past years, but Neural Networks (NNs) are nowadays the standard one.

NNs are flexible computing frameworks and universal approximators that can be applied to a wide range of learning problems with a high degree of accuracy. The main idea of NN is to extract linear combination of the inputs (in the problem considered here, Fault Detection and Classification (FDC) data from diagnostic systems) and then model the target (wafer deposition thickness) as a nonlinear function of such features. However, NNs can be really hard to train in learning problems with high dimensionality, as is the case in semiconductor manufacturing modeling. Moreover, given the use of non-linear features of the inputs during the algorithm training, the results often lack of interpretability.

A NN is a network of interconnected artificial neurons (ANs) where the outputs are weighted, possibly nonlinear transformations of the inputs. NN-based models exhibit excellent flexibility and computational properties.

A NN is composed by 3 kinds of layers:

- an *input layer* (L_{in}), where the corresponding parameters are associated with input variables (in the problem considered here, the FDC parameters);

ISBN: 978-1-61804-131-9

- a *hidden layer* (L_{hidden});
- an *output layer* (L_{out}), where the nodes correspond to the parameters that have to be predicted (thickness, as far as the CVD process is concerned).

In this paper we consider Feed-Forward NN where no loops are present between the layers. It has been shown that a Feed-Forward NN with one hidden layer can approximate any function, and this is the most used scheme amongst NN in black-box identification.

In Fig. 2 a general scheme for a Feed-Forward NN with one hidden layer is shown. Nodes represent variables while arches are associated to functions that describe interconnections between variables. In the scheme there are $\{u_i\}_{i=1}^{m_I}$ inputs and $\{w_i\}_{i=1}^{m_O}$ outputs to be modeled. Features $\{v_i\}_{i=1}^{m_H}$ are created from linear combinations of the inputs

$$v_i = h_a(\alpha_{0i} + \alpha_i^T U), \quad i = 1, \ldots, m_H, \qquad (1)$$

where U is the matrix of the inputs, while outputs in turn are created from linear combinations of the created features

$$w_i = h_b(\beta_{0i} + \beta_i^T V), \quad i = 1, \ldots, m_O, \qquad (2)$$

where V is the matrix of the hidden features. The *activation function* $h_a(\cdot)$ is usually chosen to be non-linear (sigmoid, arctan, radial-basis function[6]), while the *output function* $h_b(\cdot)$ is typically chosen linear for regression problems.

Coefficients $\alpha_\cdot$ and $\beta_\cdot$ are called *weights* and are chosen such to minimize the Mean Squared Error (MSE)

$$\text{MSE} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}, \qquad (3)$$

where, as in previous Section, n is the number of observations, y the real output and $\hat{y}$ the predicted output.

NNs are usually trained by using the *back-propagation* algorithm, according to which, weights are computed in a two-phase procedure where, after an initial guess, the prediction error are computed and then propagated backwards in the NN structure to correct the weights; then, with the new weights, a the new prediction errors are computed; this procedure is iterated several times in order to reach small values of (3). For more details we refer the reader to [2].

3 The Proposed Clustering Approach

Smart data clustering is an important element to obtain accurate VM systems. As stated before, on a semiconductor manufacturing equipment, hundreds of different products are run, each one of them with its own tool settings; if we want to model each one of these products singularly, we will have several of them with really few data, not enough to build a reliable statistical model. However a great amount of products present similar FDC data; this can be appreciated by exploiting Principal Component Analysis (PCA) for visualization.

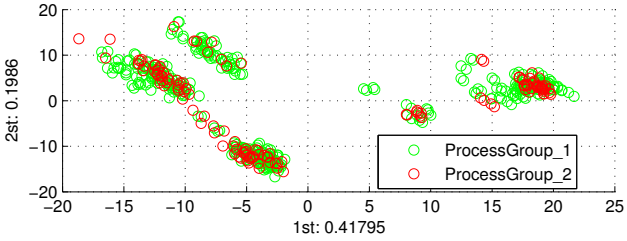

Figure 3: First two PCs for two different products (ProcessGroup_1 and ProcessGroup_2). In the axes, the variability explained by the corresponding PC is reported.

PCA is a linear projection-based method that transforms a set of uncorrelated variables into a new set of uncorrelated variables, named Principal Component (PS). PCA is applied to the input matrix X. X is written in terms of the $n \times l$ scores matrix T and the $p \times l$ loadings matrix P, plus a residual matrix E, as follows

$$X = TP^T + E \qquad (4)$$

$$= \sum_{i=1}^{l} t_i p_i^T + E \qquad (5)$$

where $t_i = Xp_i$. The vectors p_i are named Principal Components (PCs) [2]. PCs are arranged in order of magnitude, and the first PC can be geometrically interpreted as the direction where most of the variability of X is explained. Other PCs define directions where less and less variability is explained. By analyzing the magnitudes of the PCs, it is possible to employ only $l < p$ parameters to construct the model.

PCA can be employed for data distribution analysis: by visualizing the first 2/3 PCs, PCA is a useful tool to develop insights on distributions of high dimensional datasets. In Fig. 3) the first two PCs for two different products are reported; for those products whose FDC data distributions are "similar", it is reasonable to model them together in order to increase the amount of data available and consequently the confidence on the statistical model. On the other hand, it would be impossible to examine through PCA all pairs of products, and, besides, visualizing the first

ISBN: 978-1-61804-131-9

2/3 PCs could not be enough to discriminate if two products are statistically "close" or not.

We propose here a *quantitative* approach to clustering based on the statistical distance of products data distributions.

Let P and Q be the probability distributions for two different products. We define with $D_f(P\|Q)$ an f-divergence function [1] that measures the difference between P and Q

$$D_f(P\|Q) = \int_{\mathbb{R}^p} f\left(\frac{dP}{dQ}\right) dQ. \qquad (6)$$

f-divergence are non-negative, monotone, and convex functions; the most famous f-divergences are the Kullback-Leibler divergence [5] and the Hellinger distance [8], that enjoys the property of being simmetric for P and Q ($D_f(P\|Q) = D_f(Q\|P)$). The data clustering based on the statistical distance defined by (6) is illustrated in Algorithm 1.

Algorithm 1: f-divergence-based data clustering for chamber A.

Data: FDC tool data.
Result: Data Clustering.

1. For each couple of products $\{i, j\}$, with probability distributions P_i and P_j, in the production dataset we compute the $D_f(P_i\|P_j)$.
2. *First Clustering* - If $D_f(P_i\|P_j) < t$, where $t > 0$, is a "small" threshold, we group together products i and j.
3. *Second Clustering* - If a group of products, or a single product, G has, after step 2), a total amount of observations N_G that is smaller than a threshold T_n, we add to this group the data of the product i_1 outside this group with the smallest $D_f(P_{i_1}\|P_G)$. We iterate this operation adding other products i_2, i_3, ... until $N_G \geq T_n$.

To estimate distributions from data, it is possible to use a Kernel Density Estimator [9].

4 Experimental

To illustrate the proposed clustering approach, the problem of estimating wafer deposition thickness after the Chemical Vapor Deposition (CVD) process step is considered.

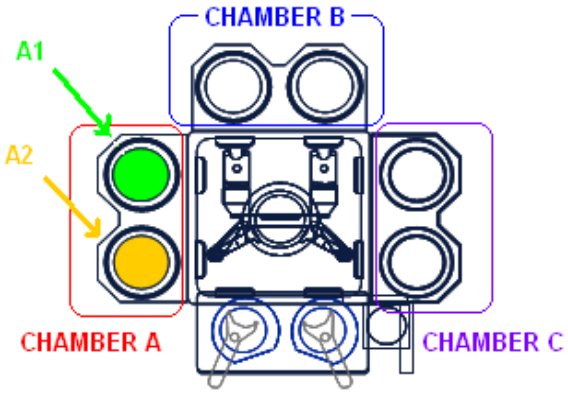

Figure 4: Scheme of a CVD equipment: each machine has 3 separated chambers (A, B and C), each one of them divided into two sub-chambers (1 and 2).

4.1 Chemical Vapor Deposition

The CVD process is used to produce a thin film over the wafer that is used as a substrate. The CD of this process is the thickness of the deposited layer. The CVD process is the first in line in the semiconductor manufacturing sequence. The processing of wafers defected in the first stages of the sequence, but not detected as such, clearly results in a waste of resources. VM systems are therefore even more crucial for this particular process than for others given the importance of monitoring wafer quality in such early production stages.

For the CVD process, measures of the deposition thickness are available for sample wafers from a lot, whereas FDC data from the tool are available for every processed wafer. Such equipment variables, together with the actual measurements, are then used as predictors in the mathematical model.

Hundreds of variables act in the CVD process and the dataset is even further complicated by the fact that each product has a different target (deposition thickness objective), the equipment is composed of 3 separated chambers (A, B and C) that exhibit different behaviors, and each chamber is divided into two sub-chambers (1 and 2) (the structure of the CVD equipment is shown in Fig. 4).

The problem of modeling the CVD process has been approached by using different techniques, both Linear, like Ordinary Least Square (OLS) and Partial Least Squares (PLS), and Non-Linear, like Artificial Neural Networks (NNs). It has been shown [3] that NNs guarantee better performance in modeling semiconductor manufacturing processes than other linear approaches; for this reason we will NNs as modeling tool for the simulation of this work.

ISBN: 978-1-61804-131-9

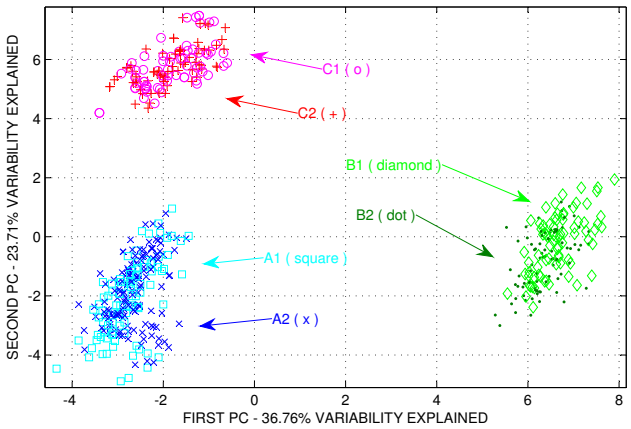

Figure 5: First two PCs different products: chambers highlighted. In the axes is reported the variability explained by the corresponding PC.

4.2 Modeling with Different Clustering Approaches

We test the clustering approach based on f-divergence on the entire production dataset available that consists of N products. The clustering approach described in Alg. 1 is done for each chamber separately; as can be appreciated in Fig. 5 by exploiting PCA each chamber can be in fact considered as a different tool.

The available dataset consists of different products with four different targets coming from the same CVD tool. We indicate with CLA, CLB and CLC the clusters of data regarding chambers A, B and C and with CL$T1$, CL$T2$, CL$T3$ and CL$T4$ the ones regarding the targets. We have considered the 10 products with more data available, for a total of $n = 6703$ data, with $n_{\text{CL}A} = 2410$, $n_{\text{CL}B} = 2205$ and $n_{\text{CL}C} = 2088$ and $n_{\text{CL}T1} = 1341$, $n_{\text{CL}T2} = 1809$, $n_{\text{CL}T3} = 1686$ and $n_{\text{CL}T4} = 1867$.

The evaluation of performance of each method is done through Repeated Random Sub-Sampling Validation, also known as *Monte Carlo crossvalidation (MCCV)*, where K simulations are done by randomly splitting the n_{CL} observations into a training dataset of $\lfloor n_{CL}q \rfloor$ maintenance cycles and a validation dataset of $\lceil n_{CL}(1-q) \rceil$ maintenance cycle, with $0 < q < 1$. It has been shown that MCCV is asymptotically consistent resulting in more pessimistic predictions of the test data compared with full crossvalidation.

For each product $i = 1, 2, \ldots, N$ we have n_i observations; we split them through MCCV into a training dataset of $\lfloor 0.7n_i \rfloor$ maintenance cycles and a validation dataset of $\lceil 0.3n_i \rceil$.

We compare several kind of clustering:

- *Chambers* - products processed in the same chamber are modeled together;
- *Target* - products with the same target CVD thickness are modeled together;
- *f-divergence* - clustering as described in Alg. 1, with different models for different chambers;
- *None* - a model for each product and for each chamber.

Clustering	MSE	MAPE
None	13.8902	8.0153
Chamber	9.4217	5.3997
Target	10.9382	6.3140
f-divergence	**6.1954**	**3.6183**

Table 1: Clustering performances: average MSE and MAPE for each cluster on $K = 1000$ simulations.

The performance in terms of MSE (3) and *Mean Absolute Percentage Error* (MAPE)

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right|. \qquad (7)$$

of the proposed clustering approaches are summarized in Table 4.2. The modeling has been done through NNs in each clustering approach.

Some comments on the experimental results reported in Table 4.2

- all kind of clusterings guarantee better performances than the "no-clustering" approach, underlying the need of a clustering step before modeling;
- chamber clustering guarantees lower MSE than target clustering, suggesting than a qualitative approach to clustering may not be as successful as a quantitative approach to input data grouping, that is therefore highly recommended;
- the f-divergence clustering outperforms all the other approaches.

5 Conclusion

A new data clustering algorithm for VM and and SS has been proposed. The new approach is based on Information-Theory elements and allows to automatically group different logistic cases in the production set that may be modeled together. This task is of paramount importance for the modeling of the entire production set in high-fragmented manufacturing processes.

Acknowledgements: This work has been done within the framework of IMPROVE (Implementing

Manufacturing science solutions to increase equiPment pROductiVity and fab pErformance), an European Nanoelectronics Initiative Advisory Council project [4]. The authors would like to thank A. Schirru and C. De Luca of INFINEON TECHNOLOGY AG, Site Villach, for providing the raw data used in this paper

References:

[1] S. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28:131–142, 1966.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer, 2009.

[3] M.-H. Hung, T.-H. Lin, F.-T. Cheng, and R.-C. Lin. A novel virtual metrology scheme for predicting cvd thickness in semiconductor manufacturing. *IEEE/ASME Transactions on Mechatronics*, 12:308–316, 2007.

[4] ENIAC IMPROVE. Official website. In *www.eniac-improve.eu*, Retrieved May 5th, 2012.

[5] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[6] Y. Lu, N. Sundararajan, and P. Saratchandran. Performance evaluation of a sequential minimal radial basis function (rbf) neural network learning algorithm. *IEEE Transactions on Neural Networks*, 9:308–318, 1998.

[7] S. Pampuri, A. Schirru, G.A. Susto, G. De nicolao, A. Beghi, and C. DeLuca. Multistep virtual metrology approaches for semiconductor manufacturing processes. In *8th IEEE International Conference on Automation Science and Engineering*, pages 91–96, 2012.

[8] D.E. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.

[9] D.W. Scott. *Multivariate Density Estimation*. Wiley Press, 1992.

[10] G. A. Susto and A Beghi. Least angle regression for semiconductor manufacturing modeling. In *IEEE Multi-Conference on Systems and Control*, 2012.

[11] G A Susto, A Beghi, and C DeLuca. A virtual metrology system for predicting cvd thickness with equipment variables and qualitative clustering. In *IEEE Conference on Emerging Technologies & Factory Automation*, 2011.

[12] G.A. Susto, A. Beghi, and C. De Luca. A predictive maintenance system for epitaxy processes based on filtering and prediction techniques. *IEEE Transactions on Semiconductor Manufacturing*, 99:(To appear), 2012.

[13] G.A. Susto, S. Pampuri, A. Schirru, and A. Beghi. Optimal tuning of epitaxy pyrometers. In *Proceeding of 23rd IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 294–299, 2012.

[14] G.A. Susto, A. Schirru, S. Pampuri, and A. Beghi. A predictive maintenance system based on regularization methods for ion-implantation. In *Proceeding of 23rd IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 175–180, 2012.

[15] G.A. Susto, A. Schirru, S. Pampuri, G. De Nicolao, and A. Beghi. An information-theory and virtual metrology-based approach to run-to-run semiconductor manufacturing control. In *8th IEEE International Conference on Automation Science and Engineering*, pages 354–359, 2012.

[16] M.T. Tham, G.A. Montague, A.J. Morris, and P.A. Lant. Soft-sensors for process estimation and inferential control. *Journal of Process Control*, 1:3–14, 1991.