Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Principal Component Analysis Based Support Vector Machine for the End Point Detection of the Metal Etch Process

Kyounghoon Han*, Seunghyok Kim, Kun Joo Park***, En Sup Yoon****, and Heeyeop Chae†**

*School of Chemical & Biological Engineering, Seoul National University, Seoul, Korea (e-mail: *mokdong@pslab.snu.ac.kr, **hyok20@pslab.snu.ac.kr, ****esyoon@pslab.snu.ac.kr)*
****Semiconductor business part, DMS Co., Ltd., Suwon, Gyeonggi-do, Korea (e-mail: kjpark@dms21.co.kr)*
†Department of Chemical Engineering, Sungkyunkwan University, Suwon, Gyeonggi-do, Korea (e-mail: hchae@skku.edu)

Abstract: An endpoint detection using the algorithm of principal component analysis based support vector machine was developed for the plasma etching process. Because many endpoint detection techniques use a few manually selected wavelengths, noise render them ineffective and it is hard to select the important wavelengths. So the principal component algorithm with the whole wavelengths has been developed for the more effective monitoring of end point. And the support vector regression was followed for the real-time end point detection with reduced wavelengths to save the processing time. This approach was applied and demonstrated for a metal etching process of Al and 0.5% Cu on the oxide stack with inductively coupled BCl_2/Cl_2 plasma.

1. INTRODUCTION

In semiconductor industry, plasma etching was developed in a variety of names – plasma etching, plasma assisted etching, reactive sputter etching, reactive ion etching, etc. These all nevertheless rely on the same basic principle – removing unwanted materials by the formation of volatile reaction productions in a glow discharge. When the target materials or layer is removed, it needs to stop the plasma etching process exactly to avoid excessive over-etching and this event is called as the end point detection (EPD).

The optical emission spectroscopy (OES) technology is the most widely used method for the simple EPD monitoring. This OES uses an optical emission spectrometer for tracing of the reactive species in plasma reaction with the removable materials. Most of the EPD methods using OES focus on identifying a single wavelength corresponding to a chemical species which shows a pronounced transition at the end point time. When the target layer is cleared by etching process, the concentration of reaction product from the target layer is reduced and the one from the under-lying is increased.

Biolsi *et al*. [1999] demonstrated an advanced endpoint system for small open-area etching by applying threshold signal processing with single wavelength. This single wavelength method cannot avoid the noise problem or time delaying associated with filtering. Furthermore selection of appropriate wavelength requires significant experience of process engineers. However, this single wavelength method shows its detection limitation when the open area (surface of target materials) is small or the signal is not strong enough. So this method usually works reliably for only for large open area wafers (typically larger than 10%). White *et al.* [2000] proposed T^2 and Q statistics for the endpoint detection of low open-area wafers using principal component analysis (PCA) in conjunction with T^2 detection and recursive mean update. This method works also reliably for large open areas (>10%) because the endpoint feature is severely corrupted with a drift in small open area. [Yue and co-workers, 2001] Yue and co-workers [2001] also used PCA with extracted a reliable endpoint signal using the loading vectors. But the reduced data can cause decreasing of sensitivity in PCA algorithm.

In this paper, we used PCA for data reduction to save the processing time and for the models of EPD monitoring. Then we estimated for real-time EPD by support vector regression for increasing its sensitivity.

2. THEORY

2.1 Principal Component Analysis

Because PCA [Jackson, 1991] and Support Vector Machine (SVM) [Vapnik, 1998] are very famous tools in chemometrics nowadays, the brief introductions were shown in this section.

PCA decomposes the data matrix (**X**) as the sum of outer product of vectors $\mathbf{t}_i$ and $\mathbf{p}_i$ plus a residual matrix (**E**):

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}^T{}_1 + \mathbf{t}_2\mathbf{p}^T{}_2 + \ldots + \mathbf{t}_k\mathbf{p}^T{}_k + \mathbf{E} \tag{1}$$

where k must be less than or equal to the smaller dimension of **X.** The $\mathbf{t}_i$ vectors are defined as ***scores,*** and contain information on how the samples relate to each other. The $\mathbf{p}_i$

vectors are known as ***loadings*** and contain information on how the variables relate to each other. In the PCA decomposition, the $\mathbf{p}_i$ vectors are the eigenvectors of the covariance matrix, *i.e.* for each $\mathbf{p}_i$:

$$\mathrm{Cov}(\mathbf{X})\mathrm{p}_i = \lambda_i \mathbf{p}_i \tag{2}$$

where the λ_i is the eigenvalue associated with the $\mathbf{p}_i$. Note that for $\mathbf{X}$ and any $\mathbf{t}_i$, $\mathbf{p}_i$ pair:

$$\mathbf{X}\mathbf{p}_i = \mathbf{t}_i \tag{3}$$

The score vector ($\mathbf{t}_i$), is the linear combination of the original $\mathbf{X}$ variables defined by $\mathbf{p}_i$. Because the score vector needs normalization of data the concept of 'product' was used for real-time monitoring of score vector without normalization.

$$\mathbf{Y}_i = \mathbf{X'}\,\mathbf{P}_i \tag{4}$$

$\mathbf{Y}_i$ means product value of ith sample time and $\mathbf{X'}$ means the data matrix without normalization. The endpoint can be decided by the monitoring of the estimated product with raw data of real time wafer and loading vector of model wafer. The ratio of two products can be used for EPD to increase the sensitivity of monitoring after comparison of all products values.

2.2 Support Vector Machine for Regression

The SVM is a learning system that uses a hypothesis space of linear function in a high dimensional feature space, which is trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. [Vapnik, 1995; Cherkassky and Muler, 1998]

Suppose there is a set of training data $\{(x_1,y_1), \ldots, (x_n,y_n)\} \subset \mathbf{X} \times \Re$, where $\mathbf{X}$ denotes the space of the input patterns. The SVM considers approximating functions of the form where the kernel function $\phi_i\ (x)$ are features, as a nonlinear classification.

$$f(x,w) = \sum_{i=1}^{n} w_i \phi_i(x) + b \tag{5}$$

where $\phi_i\ (x)$ represents a mapping from the input space (x) to a feature space f and b also represents a bias term. A more generalized form for SVM uses a kernel function $K(x_i, x_j)$ is which is the inner production of point $\phi_i\ (x_i)$, $\phi_i\ (x_j)$ mapped into feature space. The use of kernels makes it possible to map the data implicitly a feature space and to train a linear machine in such a space, potentially side-stepping the computational problems inherent in evaluating the feature map. A list of popular kernels is shown in Table 1.

Table 1. Different types of kernel function

Name of kernel	Expression
Polynomial of degree p	$K(x_i,x_j) = [(x_i,x_j)+1]^p$
Gaussian RBF	$K(x_i,x_j) = \exp\left(\frac{\lVert x_i - x_j \rVert^2}{-2\sigma^2}\right)$
Multilayer perceptron	$K(x_i,x_j) = \tanh[(x_i,x_j)+b]$

Vapnik [1995] introduced a general type of error function, which is known as the linear loss function with ε-insensitivity zone.

$$|y - f(x,w)|_\varepsilon = \begin{cases} 0; & if\ |y - f(x,w)| \le \varepsilon \\ \left||y - f(x,w) - \varepsilon;\right| & otherwise \end{cases} \tag{6}$$

The loss is equal to zero if the difference between the predicted value of $f(x, w)$ and the measured value is less than ε. The support vector regression needs to minimize the following risk function.

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{7}$$

$$\text{subject to} \quad \begin{Bmatrix} y_i - w\phi(x_i) - b_i \le \varepsilon + \xi_i^* \\ w\phi(x_i) + b_i - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{Bmatrix} \tag{8}$$

where ξ and ξ^* are slack variables, which have positive values in order to quantify the non-separable data in the defining condition of the hyperplane. The constant, C can be determined by the trade off between the model complexity of f and its accuracy on the training data. The parameters used in support vector regression are shown in Fig.1. This constrained optimization is solved by forming a primal variables Lagrangian, $L_p(w,\ \xi,\ \xi^*)$

$$L_p(w, b, \xi, \xi^*, \alpha_i, \alpha_i^*, \beta_i, \beta_i^*) =$$

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i^*(y_i - w\phi(x_i) - b - \varepsilon + \xi_i^*)$$
$$-\sum_{i=1}^{n}\alpha_i(w\phi(x_i) + b - y_i + \varepsilon + \xi_i) - \sum_{i=1}^{n}(\beta_i^*\xi_i^* + \beta_i\xi_i) \tag{9}$$

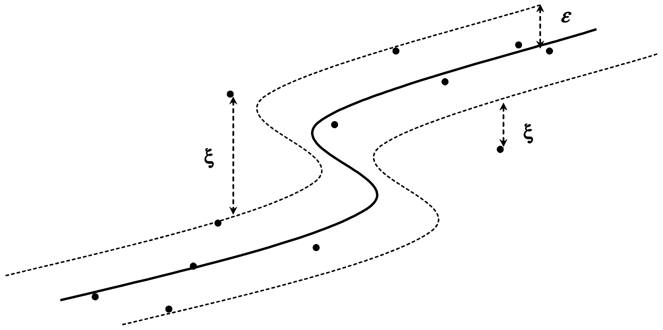

Fig. 1. The insensitive band for a non-linear regression function [Vapnik, 1995]

Lagrangian $L_p(w, b, \xi, \xi^*, \alpha_i, \alpha_i^*, \beta_i, \beta_i^*)$ must be minimized with respect to the primal variables, w, b, ξ and ξ^*, and maximized with respect to non-negative Lagrange multipliers α, α^*, β, and β^*. Again, this problem can be solved either in a dual space. By applying the Karush–Khun–Tucker (KKT) conditions for regression, the dual variables Lagrangian $L_d(\alpha, \alpha^*)$ are maximized;

$$L_d(\alpha,\alpha^*)$$

$$= -\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)y_i-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j) \quad (10)$$

$$\text{subject to} \quad \sum_{i=1}^{n}\alpha_i^* = \sum_{i=1}^{n}\alpha_i \quad (11)$$

$$0 \le \alpha_i, \alpha_i^* \le C \qquad \text{i=1,…,n}$$

With assumption of ξ^*, in the range of $0 < \alpha_i^*, < C$ the regression represents using the previous kernel function as followings:

$$f(x) = \sum_{i=1}^{n}(\alpha^*_i - \alpha_i)K(x_i, x) + b \quad (12)$$

3. PCA BASED SVM FOR EPD

3.1 Principal component analysis for the 'product' with the whole wavelengths

Initially, the entire range of OES signals from the standard wafer was captured and normalized. The covariance of this normalized data was obtained and singular value decomposition (SVD) performed. The loading vectors were obtained from solving of eigenvalue problem of the result of its SVD. Finally, the product value of the standard wafer was calculated from the multiplication of raw OES data and loading vectors as shown in Fig. 2.

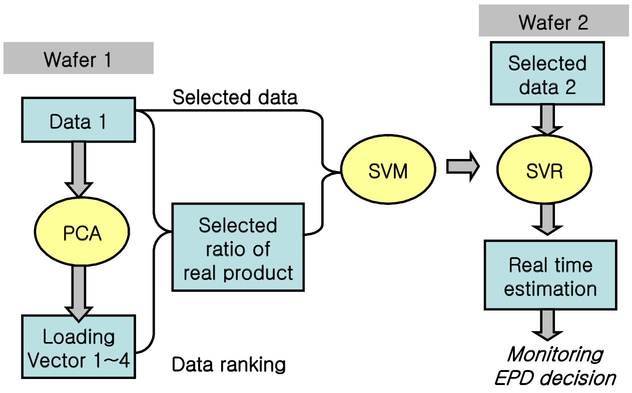

Fig. 2. PCA based SVM model for real-time EPD

3.2 PCA based wavelengths reduction

If you should use and save the whole OES data, there can be magnificent memory burdens for the real-time processes. The reduction technique for the wavelengths with important information was introduced by Yue et al [2001]. The wavelengths from the standard wafer can be sorted by using the loading vectors for their importance.

$$l_i = \sqrt{\sum_{j=1}^{n} l_{ij}^{\,2}} \quad (13)$$

where l_{ij} denotes the loading vectors for the i^{th} wavelength and j^{th} PC. In general, two or three loading vectors were selected, which have the main information over than 80%. And the reduced wavelengths of the real time target wafer can be selected by above criteria.

3.3 SVR with reduced wavelengths for EPD

The SVM was learned with reduced wavelengths as input variables and with the selected product ratio of whole wavelengths as output variables in the standard wafer. Then the reduced wavelengths from the real-time target wafer can estimate the product ratios by support vector regression (SVR). EPD can be decided by the monitoring of the product ratio values, which change significantly. This standard wafer should be updated periodically to consider the process drift as shown in Fig. 3. The more rapid the standard wafer can be updated, the more exact the EPD prediction can be exact. Because the vast data change can cause the process memory failure, the period of model update should be optimized.

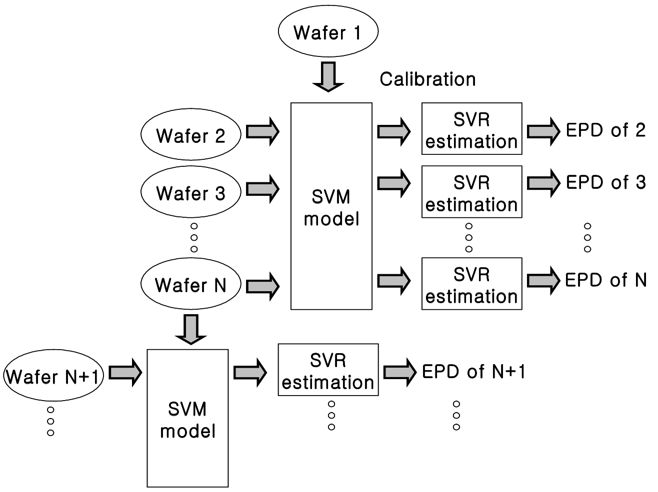

Fig. 3. Real-time EPD estimation model update

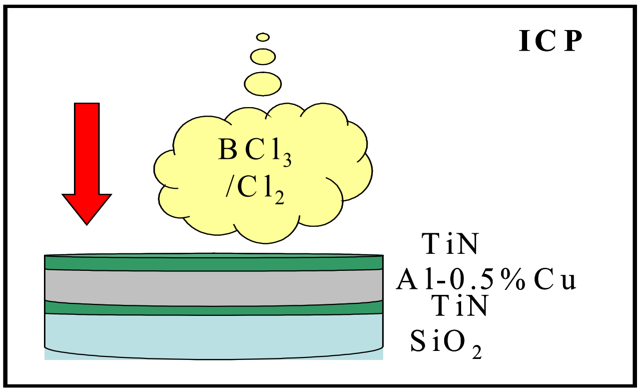

Fig. 4. Wafer composition for the metal etch [Wise, B. M, *et al.*, 2004]

4. CASE STUDY

For the case study, the open data source of the monitoring problem in the semiconductor processing was used. (http://www.software.eigenvector.com/Data/Etch/, Eigenvector Research Inc.) The goal of this process was to etch a TiN/Al-0.5%Cu/TiN/oxide stack to form a metal line employing an inductively coupled BCl_3/Cl_2 plasma, as shown in Fig. 4. Our focus was only on an Al-stack etch process, which was performed on the commercially available Lam 9600® plasma etch tool. The OES data were collected, which consist of 40 process set points and the measured variables sampled at 1 second intervals. This experiment was done at three different times considering the condition drift, and consisted of 126 wafers. These experiments were divided into the 3 different experiment groups for 3 different experiment times. In the first experiment, the 1, 11, 21, 31 wafers were used for model wafer and EPDs of another wafers were predicted by our algorithm. And the model was updated for every 10 wafers. Similarly there were four modelling groups in second experiment, and third experiment for prediction

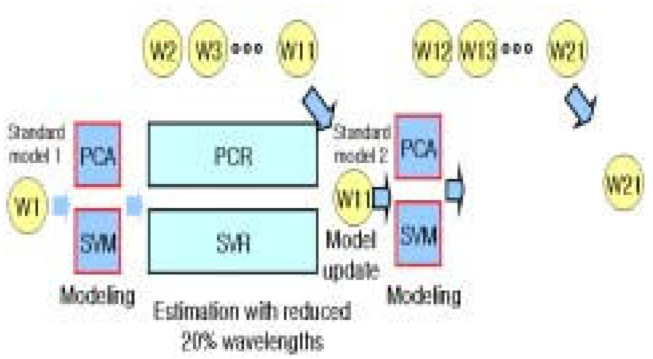

Fig. 5. Real-time SVM model update

and model update. Fig. 5 shows the real-time model update for the SVM prediction for EPD.

5. RESULTS

The optical emission signals of the 129 channels from 250nm to 791.5nm were measured. The normalization processed for PCA and singular value decomposition (SVD) was performed with the covariance of these normalized data of each model wafer. Then, the loading vectors were obtained from the solving of eigenvalue problem of this result of SVD. The entire range of OES signals was multiplied with the loading vector for drawing the product line.

Fig. 6 shows the 4-dimensional data set which was composed of the signal intensity (arbitrary unit), samples (sec), channels (nm), batch (waver number). These 129 channels of theses model wafer were ranked by the PCA based wavelengths reduction algorithm. The SVM was used for each model with reduced high ranked 20% channels as input value and previous 2 product line (first, second) as output value. The kernel function used for modelling was the polynomial with 2^{nd} degree and the parameter C was set to 5,000. The loss function with ε-insensitivity zone was wet to zero for the simple regression in this study. Then, we monitored the ratio of the predicted first product value over the second value. Finally the SVR was used for the real-time EPD prediction with real-time 20% reduced channel data (26 channels). Because the first wafer cannot be monitored at each three experiment, the 123 wafers from the 126 wafers were monitored for EPD using both single PCA and PCA-based SVM method.

Table 2. Wafer classification

Experiment		Wafer numbers			
1	Model	*1*	*11*	*21*	*31*
	Target	2-11	12-21	22-31	32-42
2	Model	*43*	*53*	*64*	*73*
	Target	44-53	54-63	64-73	74-85
3	Model	*86*	*96*	*106*	*116*
	Target	87-96	97-106	107-116	117-126

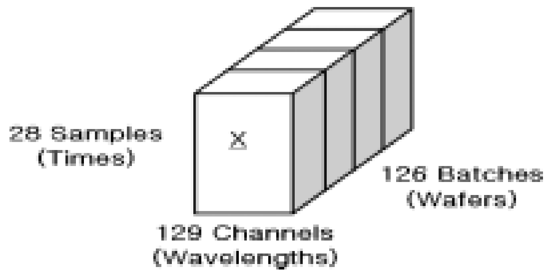

Fig. 6. Data stack of metal etch with plasma gas

Fig. 7, 8, 9, 10 shows the monitoring results proposed by our algorithm (up) and single PCA estimation in several important wafers (down). We could find the exact change of the prediction curve about these times using our algorithm about 13 to 20 seconds whereas there were some fluctuations using the single PCA prediction with the small data in figure 7 (29th wafer), 8 (33th wafer), 9 (58th wafer), 10 (88th wafer).

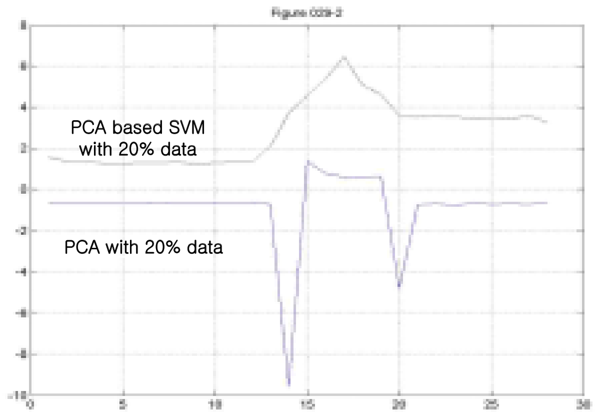

Fig. 7. EPD monitoring comparison of the 29th wafer

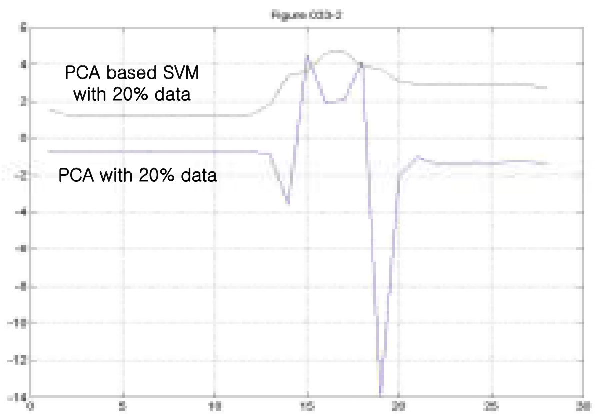

Fig. 8. EPD monitoring comparison of the 33th wafer

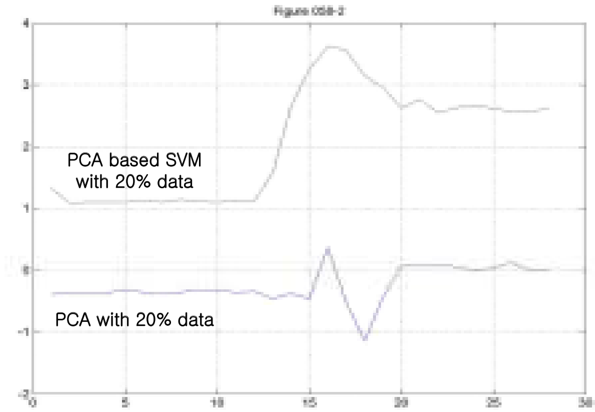

Fig. 9. EPD monitoring comparison of the 58th wafer

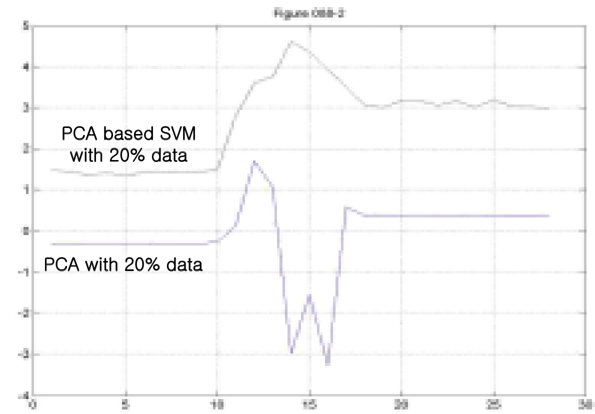

Fig. 10. EPD monitoring comparison of the 88th wafer

The monitoring results of 120 wafers (98%) showed exact end point behaviours using PCA-based SVM method except 3 wafers because of faulty conditions. But only 59 wafers (48%) showed the end point behaviours using single PCA method as shown as the bad cases in Fig. 7 to Fig. 10.

6. CONCLUSIONS

In this paper, the PCA based SVM algorithm was developed for the real-time end point prediction of the plasma etching process. This algorithm was applied to the Al-0.5%Cu metal etch stack with BCl_3/Cl_2 plasma gas in an inductively coupled ether. At first, 126 data sets were collected with 129 wavelengths per 1 second. And the PCA loading vector was calculated for the wavelengths ranking and product value from the model wafer. Then the SVM modelling was achieved with the reduced 20% high ranking data as input value and product with whole wavelengths as output value. Finally the SVR was done with the reduced wavelengths of every wafer for the real-time end point prediction. While the traditional single PCA method can show its sensitivity limitation with reduced wavelengths, the suggested PCA based SVM model shows better prediction figures as a result with small process burdens.

NOMENCLATURES

X: data matrix (PCA), input space (SVM)
y: output space
n: dimension
t: score
p: loading
E: residual matrix
Cov(X): covariance of matrix X
$\Re$: real number domain
C: constant of support vector regression model
$K(x_i, x_j)$: Kernel function
ε : error probability
Y: product value
f(x, w): feature space of data x with weight vector w
w: weight vector
ϕ: mapping
b: bias
ξ : margin slack variable
α_i, β_i: Lagrange multipliers
L_p: primal variables Lagrangian function
L_d: dual variables Lagrangian function
$|\xi|_\varepsilon$: ε -intensivity loss function

REFERENCES

Biolsi, P., Morvay, D., Drachinik, and Eliinger, S., "An advanced Endpoint Solution <1% Open Area Applications; Contact and Via", IEEE/SEMI Advanced Semiconductor Manufacturing Conference, 391 (1996)

Chiang, L. H., Russell, E. L. and Braatz, R. D., *Fault Detection and Diagnosis in Industrial Systems*, Springer-Verlag (2001)

Dreeskornfeld, L., et al. "Reactive ion etching end point detection of microstructured Mo/Si multilayers by optical emission spectroscopy", *Microelectronic engineering*, **54**, 303 (2000)

Grill, A., *Cold Plasma in Materials Fabrication*, IEEE Press (1994)

Kim, H. S., et al., "Etch end-point detection of GaN-based device using optical emission spectroscopy", *Materials Science & Engineering*, **B82**, 159 (2001)

Kim K. S. and Ko J. W., "Real-time risk monitoring system for chemical plants", *Korean J. Chem. Eng*. **22(1)**, 26 (2005)

Pearton, S. J., et al., "Optical emission end point detection for via hole etching in InP and GaAs power device structures", *Materials Science Engineering*, **B23**, 36 (1999)

Qin, S. J., et al, "Semiconductor manufacturing process control and monitoring : A fab-wide freamwork", *Journal of Process Control*, **16**, 179 (2006)

Seo, S. T., et al., "Run-tu-run control of inductively coupled C2F6 plasma of SiO_2 : Multivariable controller design and numerical application.", *Korean J. Chem. Eng*. **23(2)**, 199 (2006)

Vladimir N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, Inc. (1998)

White, D., Goodlin, B. E., Gower, A. E., Boning, D. S., Chen, H. Sawin, H. H. and Dalton T. J., "Low open-area endpoint detection using a PCA-based T2 statistic and Q statistic on optical emission spectroscopy measurements", *IEEE Trans. Semiconductor Manufacturing*, **13(2)**, 193 (2000)

Wise, B. M., et al., "A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in semiconductor etch process", *Journal of Chemometrics*, **13**, 379 (1999)

Wise, B. M. et al., *PLS_Toolbox 3.5 for use with MATLAB™*, Eigenvector Research (2004)

Yue. H. H., et al., "Plasma etching endpoint detection using multiple wavelengths for small open-area wafers", *Journal of Vacuum Science & Technology*, **A 19**, 66 (2001)

http://www.software.eigenvector.com/Data/Etch/