Support Vector Machine-Regression을 이용한 주기신호의 이상탐지

박승환·김준석·박정술·김성식·백준걸†

고려대학교 산업경영공학과

A Fault Detection of Cyclic Signals Using Support Vector Machine-Regression

SeungHwan Park·JunSeok Kim·Cheong-Sool Park·Sung-Shick Kim·Jun-Geol Baek†

Department of Industrial Management Engineering, Korea University

Key Words : Fault Detection, Cyclic Signals, Support Vector Machine-Regression

Abstract

This paper presents a non-linear control chart based on support vector machine regression (SVM-R) to improve the accuracy of fault detection of cyclic signals. The proposed algorithm consists of the following two steps. First, the center line of the control chart is constructed by using SVM-R. Second, we calculate control limits by variances that are estimated by perpendicular and normal line of the center line. For performance evaluation, we apply proposed algorithm to the industrial data of the chemical vapor deposition process which is one of the semiconductor processes. The proposed method has better fault detection performance than other existing method

1. 서 론

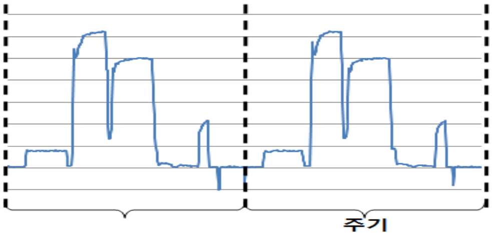
최근 전자제품의 수요가 급격히 증가함에 따라 반도체, LCD 시장은 점점 더 커지고 있다. 시장이 확대되고 있고 이에 따라 반도체 회사들은 고부가가치 창출 및 전략적 생산량의 증대를 위해 노력하고 있다. 고부가가치 창출을 위해서는 고품질, 고집적화된 제품을 생산하는 것이 중요하다. 이를 위해 기업들은 보다 미세하고 정밀한 제품을 생산하기 위해 진보된 제조 설비들을 도입하고 있다. 제조 설비의 발달에 따라 보다 정확한 공정 제어 및 관리 기술의 필요성이 증대되고 있다. ITRS (International Techonology Roadmap for Semiconductors)가 제시한 국제 반도체 기술 로드맵을 살펴보면, 45nm이하의 공정에 대한 공정 제어 및 관리 기술은 미비한 실정이다[9]. 따라서 고도화, 미세화된 공정에서 고품질, 고집적화된 제품을 생산하기 위해서는 보다 정밀한 공정 제어 및 관리 기술이 필요하다.

본 연구는 공정 제어 및 관리 기술 중 이상탐지 기법(Fault Detection : FD)을 대상으로 한다. 이상탐지는 공정 중 발생하는 온도, 압력, 가스의 농도 등의 데이터의 관찰을 통해 이루어진다. 또한 관찰된 데이터를 판단하여 공정이 정상적으로 진행되고 있는지 아닌지를 검증한다. 만일 데이터가 이상(fault)으로 판단되면 경보(alarm)를 울리거나 공정을 중지(interlock)하는 방법들을 통해 엔지니어가 기계를 정비하여 불량제품의 생산을 줄이도록 한다.

반도체, LCD 제조 공정은 반복되는 여러 단위 공정으로 구성되어 있다. 단위 공정에서는 센서(sensor)를 통해 주기별로 일정한 패턴(pattern)을 형성하는 주기신호가 발생된다. 주기신호란 한 주기가 여러 개의 구간으로 구분되어 각 구간에서 동일한 작업이 진행되는 신호를 의미한다. <그림 1>과 같이 주기신호는 매 구간

† 교신저자 jungeol@korea.ac.kr

마다 같은 패턴을 형성한다. 따라서 주기신호의 효율적인 분석을 통해 이상을 탐지하고, 공정의 불량률을 줄일 수 있는 방법이 필요하다. 본 연구는 이와 같은 주기신호를 대상으로 하는 이상탐지 방법을 제안한다.



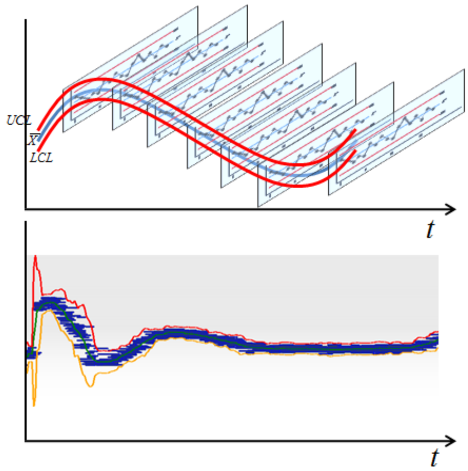
<그림 1> 주기신호 데이터의 예

주기신호 데이터의 이상탐지에 대한 기존 연구로써 전통적인 Shewhart 관리도를 사용하는 방법이 있다. Shewhart 관리도는 각 시점을 변수로 보고 해당시점의 데이터들의 통계량인 평균과 표준편차를 사용해서 중심선(Center Line), 관리 상한선(Upper Control Limit)과 관리 하한선(Lower Control Limit)을 구축한다「Montgomery, 2001」. 하지만 두 가지 문제점이 존재한다. 첫째, Shewhart 관리도는 공정에서 측정되는 주기신호에 대해 고정된 평균과 균일한 분산을 갖는 정규분포(Normal Distribution)를 따른다고 가정한다. 그러나 주기신호는 시간에 따라 이동하는 평균과 균일하지 않는 분산을 가지며 특정할 수 없는 분포를 따른다는 문제점이 있다. 둘째, 주기신호를 Shewhart 관리도에 적용하기 위해 주기신호의 구간을 나누는 방법이 제안되었으나, 구간의 길이가 다른 신호의 처리문제 및 신호 값의 변동이 큰 구간에서 분산이 증가하는 문제를 가지고 있다. 그 밖에 널리 사용되는 방법은 다변량 통계 분석 기법인 Hotelling's T2와 PCA(Principal Component Analysis), PLS(Partial Least Squares)등이 있다.「Shi and Jin, 2000」 PCA나 PLS는 잠재변수를 산출하여 차원을 축소하고 잠재변수를 이용하여 분석하는 방법이다. 잠재변수들은 연관성을 갖는 변수들의 조합으로 구성되기 때문에 이상이 발생했을 경우, 실제 원인이 되는 변수를 찾는 과정이 복잡하다. 또한, 주기신호의 길이가 각기 다르기 때문에 적용하기 어렵고, 샘플의 수보다 데이터의 길이가 크면 공분산을 계산할 수 없기 때문에 적용하기 힘들다.

본 연구는 기존의 Shewhart 관리도를 사용한 주기신호 관리방법과 제안하는 SVM-R 관리도를 비교하였다. 주기신호 관리를 위하여 반도체 공정에서 사용되던 기존의 관리도는 전체 주기신호에서 매 시점 별로 Individual 관리도를 구축하여 각기 다른 값을 갖는 중심선과 관리한계선을 연결하여 비선형의 주기신호에 대한 관리도를 구축한다. 다음은 Individual 관리도에 대한 식이다. 식(1)의 $C.V.$(critical value)는 주어진 유의수준(significant level)에 상응하는 값이다.

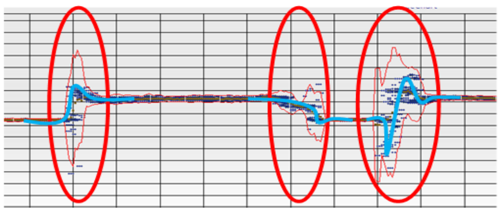
$$\begin{aligned} CL_t &= \overline{x_t} \\ UCL_t &= \overline{x_t} + C.V. \cdot s_t \\ LCL_t &= \overline{x_t} - C.V. \cdot s_t \\ (t &= 1,2,3,\ldots,n) \end{aligned} \tag{1}$$

또한 식(1)의 Individual 관리도를 종합적인 관점에서 보면 <그림 2>의 형태이기 때문에 이상이라고 판단한 Individual 관리도의 개수가 전체 주기신호관점에서 봤을 때 한계선을 벗어난 이상의 개수라고 할 수 있다.



<그림 2> 주기신호 관리를 위한 기존의 방법

하지만 기존의 관리도는 신호의 변동이 큰 구간에서 급격한 분산 증가 현상이 발생하여 관리한계선이 급격히 증폭함으로써 이상의 형태가 잘 관측될 것으로 예상되는 변동 구간에 있어서 탐지성능이 저하되는 문제가 발생한다.<그림 3>



<그림 3> 기존 관리도의 문제점

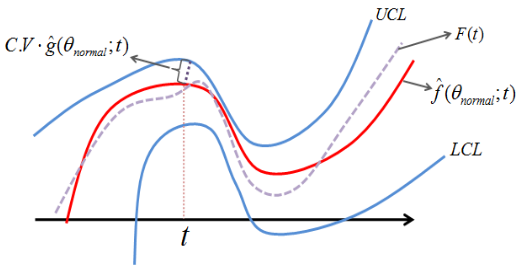
따라서 본 연구에서 제안하는 SVM-R 관리도는 비선형 형태의 주기신호를 모델링하여 관리도를 구축하였다. 따라서 기존 관리도에서 발생하는 문제점을 해결하였다.

2. SVM-R을 이용한 주기신호의 이상탐지

본 연구는 반복적으로 공정이 진행됨에 따라 얻어지는 데이터들로부터 평균을 나타내는 함수를 추정한다. $F(t)$는 공정 중 발생하는 실제 주기신호이며 다음과 같이 표현된다.

$$F(t)=f(\theta;t)+\epsilon_t,\quad \epsilon_t\sim N(0,g(\theta;t)^2) \qquad (2)$$

$F(t)$는 공정 조건 또는 환경 변화 상태(state)를 나타내는 θ와 시간 t를 파라미터(parameter)로 갖는 신호 $f(\theta;t)$와 ϵ_t의 합 으로 나타낸다. 시간 t는 매 주기에서 각 시점을 나타내고, 1,2, ...,n으로 규정되어 진다. 또한 n은 신호에 따라 임의로 결정된다. <그림 4>에서 일정 시점 t에서 관측되는 신호를 $F(t)$라고 하면, $\hat{f}(\theta_{normal};t)$는 정상조건일 때의 신호를 추정한 함수이다. 또한 ϵ_t는 백색 잡음 과정을 따르며, $\hat{g}(\theta_{normal};t)$는 정상조건일 때 ϵ_t의 표준편차를 추정하는 함수이다. 따라서 공정에서의 정상조건 θ_{normal}에 대해 $\hat{f}(\theta_{normal};t)$를 추정하고 관리도의 중심선을 찾는다. 그리고 $\hat{g}(\theta_{normal};t)$로 관리한계선을 설정하여 <그림 4>와 같은 관리도를 구축한다.



<그림 4> 가정한 주기신호의 모델

따라서 비정상 조건인 $\theta_{abnormal}$에서 발생되는 신호와 같은 이상을 탐지할 수 있다.

2.1 SVM-R을 이용한 $f(\theta;t)$ 추정

SVM(Support Vector Machine)은 지도적 학습방법(supervised learning)을 통해 여유 거리(margin)를 최대화 하는 초평면(hyperplane)을 구성하여 데이터를 분류하는 기법이다. SVM은 폭 넓은 분야에서 기존의 다른 방법 보다 뛰어난 능력을 발휘하고 있다[5,7,8]. SVM은 분류(classification)문제를 해결하기 위해 Vapnik에 의해 개발되었지만 최근에는 회귀분석이나 확률 밀도 예측(probability density estimation)과 관련된 문제를 해결하기 위해 확장되었다「Vapnik, 1995」. 또한 임의의 실수 값을 예측 할 수 있도록 SVM을 일반화한 SVM-R이 제안되었다. SVM-R은 식(3)와 같이 다차원 공간 R^M과 일차원 공간 R의 내적 공간 안에 학습데이터 집합이 주어진다. 학습데이터안의 관측값 x_i에 대해 실제로 얻어진 목표값 y_i로부터 ε만큼의 범위를 갖는 함수 $f(x)$를 찾는다. 여기서 $f(x)$는 추정하고자 하는 실제 주기신호와 같은 함수를 의미한다.

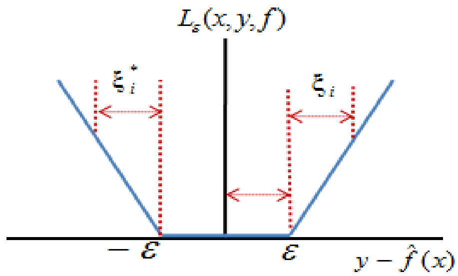
$$(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\in R^M\bullet R \qquad (3)$$

SVM-R에서 목표하는 예측함수 $f(x)$는 다음과 같은 선형(linear)모델이 된다.

$$\hat{f}(x)=\ll w,\phi(x)\gg+b, w\in R^M, b\in R \qquad (4)$$

식 (4)에서 $\ll,\gg$은 공간 R^M안에서의 벡터의 내적(dot product), w는 가중치(weight) 벡터, $\phi(x)$는 비선형 매핑 함수, b는 절편(bias)을 나타낸다. 또한 SVM-R의 목적함수는 식(5)와 같이 ε-무감도 손실함수(ϵ-insensitive loss function) $L_\epsilon(x,y,f)$를 사용하고 그림으로 나타내면 <그림 5>과 같다.

$$L_\epsilon(x,y,f)=\max(0,|y-(\ll w,\phi(x)\gg+b|-\epsilon) \qquad (5)$$



<그림 5> ε-무감도 손실함수

ϵ-무감도 손실함수는 목표값 y와 예측된 값 $\hat{f}(x)$간의 오차가 ε이하면 $L_\epsilon(x,y,f)=0$ 이고 오차가 ε보다 크면 오차의 절대값에서 ε만큼 차감한 값인 $L_\epsilon(x,y,f)=|y-\hat{f}(x)|-\epsilon$ 을 갖게된다.「박찬규, 2006)」

즉, SVM-R에서는 목표값 y_i와 예측값 $\hat{f}(x)$의 차이를 가능한 한 ε이내로 유지하면서 마진을 최대화하고, 여유거리 ξ_i와 ξ_i^*를 최소화 한다. 이를 식으로 표현하면 식 (6),(7)와 같다.

$$\min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{6}$$

$$\begin{aligned} s.t.\ \ & y_i-(w,\phi(x_i))-b \le \epsilon+\xi_i^* \\ & (w,\phi(x_i))+b-y_i \le \epsilon+\xi_i \\ & \xi_i,\xi_i^* \ge 0 \end{aligned} \tag{7}$$

식 (6),(7)를 풀기 위해서 최적 제약 조건의 학습 데이터에 대한 선형조합을 계산하기 위해 식 (8)과 같이 w를 유도한다.

$$w=\sum_{i=1}^{N}(\alpha_i-\alpha_i')\phi(x_i) \tag{8}$$

본 연구는 SVM-R에서 목표하는 $f(x)$를 비선형 형태로 확장하기 위해 커널함수(kernel function) $k(x_i,x_i')$를 사용하여 다음과 같이 구성한다.

$$f(x)=\sum_{i=1}^{N}(\alpha_i-\alpha_i')k(x_i,x_i')+b \tag{9}$$

식 (9)에서 α_i,α_i' 는 라그랑지(Lagrange) 승수이고, $k(x_i,x_i')$는 앞서 언급한 $f(x)$를 비선형 형태로 확장하기 위한 커널함수이다. 커널함수는 다음과 같은 형태로 정의 된다.

$$k(x_i,x_i')=\ll \phi(x_i),\phi(x_i') \gg \tag{10}$$

SVM-R에 사용되는 커널함수의 종류에는 Splines, Polynomial, Hyperbolic Tangent, Sigmoid, RBF (Radial Basis Function) 등이 있다. 본 연구는 식 (11)의 RBF 함수와 식(12)의 sigmoid 함수를 사용한다.「Burges, 1998」.

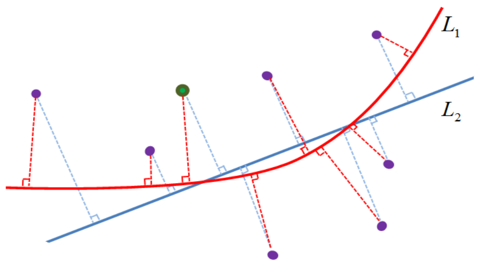
$$k(x_i,x_i')=\exp(-\|x_i-x_i'\|^2/2\sigma^2) \tag{11}$$

$$k(x_i,x_i')=\tanh(\kappa x_i-x_i'-\delta) \tag{12}$$

따라서 이상의 SVM-R 기법을 사용하여 실제 공정 주기신호 $f(\theta;t)$를 추정 할 수 있다.

2.2 분산에 대한 함수 $\{g(\theta;t)\}^2$ 추정

SVM-R을 사용해서 비선형의 주기신호를 추정했다면 주기신호의 각 시점에 대한 관리한계선 구축이 필요하다. 관리한계선을 구축하기 위해선 각 시점의 분산이 고려되어야 하고, 분산은 잔차함수 $\{g(\theta;t)\}^2$로 나타낸다. 일반적인 선형 회귀분석에서는 잔차를 구할 때 <그림 6>의 선형형태를 갖는 L_2와의 수선을 사용한다. 따라서 모든 데이터들에 대해 일정한 사영 행렬(Projection Matrix)이 요구된다. 반면 비선형 회귀분석 경우에는 <그림 6>의 비선형 형태를 갖는 L_1에 대한 수선을 사용하기 때문에, 각각 다른 사영 행렬를 고려한 잔차 계산 방법이 요구된다. 따라서 본 연구는 비선형 주기신호를 추정하므로 L_1에 해당하는 잔차를 사용한다.

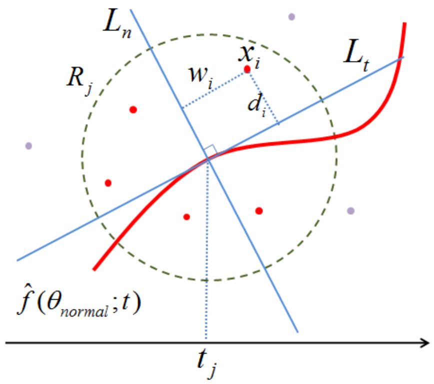


<그림 6> 선형과 비선형 회귀분석의 사영(projection) 형태

분산에 대한 함수 $\{g(\theta;t)\}^2$는 각 시점에서 접선의 기울기를 통하여 기울기에 수직이 되는 선을 기준으로 하는 데이터의 분포를 고려하여 계산한다.

<그림 7>에서 R_j는 시점 t_j에서의 분산을 고려하기 위해 시점 t_j값에 영향을 미칠 것으로 가정하는 데이터들을 포함하고 있는 영역을 나타낸다.

원 R_j안의 점들은 주기신호 데이터를 의미하고, 시점 t_j에서의 $\hat{g}(\theta_{normal};t_j)$값에 영향을 미친다. 따라서 해당 시점 $(t_j, \hat{f}(\theta_{normal};t_j))$으로부터 동일한 영역 안에 있는 데이터들로부터 분산을 추정하기 위해 범위로써 원 R_j가 필요하다.



<그림 7> 각 시점에 대한 분산 추정 모델

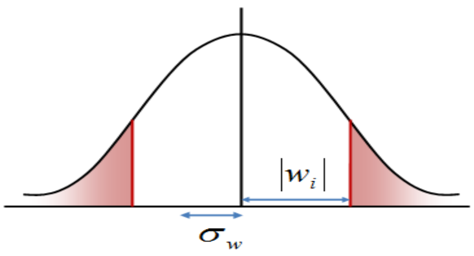
$\hat{f}(\theta_{normal};t)$의 접선 L_t와 법선 L_n을 기준으로 원 R_j 내에 속하는 데이터 x_i에 대해 수리적 방법(Numerical Method)으로 연산을 수행한다. 다음 식은 시점 t_j에서의 접선 L_t의 기울기를 나타내는 식이다.

$$slope_j = \hat{f}(\theta;t_{j+1}) - \hat{f}(\theta;t_{j-1}) / (t_{j+1} - t_{j-1})$$
$$(t_0 = 0,\ t_{j+1} = t_j + \Delta t,\ as\ \Delta t \rightarrow 0) \quad (13)$$

식(14)은 <그림 7>의 원 R_j내의 데이터 x_i 가 시점 t_j에서 $\hat{f}(\theta_{normal};t)$의 잔차의 원소일 확률을 가중치 값으로 정의한다.

$$weight_i = p(x_i \in \hat{f}(\theta_{normal};t_j)'s\ residual) \quad (14)$$

앞서 언급한 가중치를 계산하기 위해서는 정규분포함수의 확률값을 사용하고 <그림 9>는 평균이 0이고, 표준편차는 σ_w를 따르는 정규분포함수의 PDF(Probability Denstiy Function)이다.



<그림 8> 정규분포함수의 PDF

식 (15)에서 w_i는 <그림 7>에서 원 R_j내의 데이터들로부터 법선 L_n와의 거리를 나타내며 σ_w는 동일한 범위 안의 데이터들에 대한 표준편차를 의미한다. 따라서 식 (15)에서 확률값은 <그림 8>처럼 함수의 확률분포를 사용해서 구해지고, 해당 데이터 x_i에 대한 가중치로써 사용한다.

$$weight_i = \{1 - \Phi(|w_i|/\sigma_w)\} \times 2 \quad (15)$$

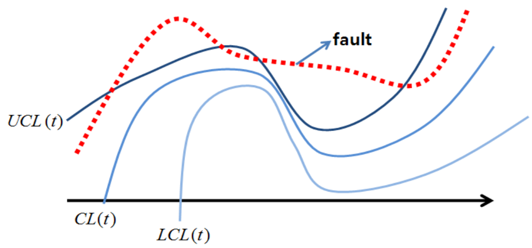
<그림 7>의 접선 L_t로부터 원 R_j에 속하는 데이터들까지의 직선거리를 나타내는 d_i와 식 (15)에서 구한 해당 시점 t_j의 각 데이터 x_i에 대한 가중치를 사용해서 정상인 상태의 주기신호에 대한 표준편차 함수 $\hat{g}(\theta_{normal};t)$를 식 (16)로 계산한다.

$$\hat{g}(\theta_{normal};t_j) \cong \sqrt{\sum_{x_i \in R_j} weight_i \times d_i^2 / \sum_{x_i \in R_j} weight_i} \quad (16)$$

최종적으로 식 (16)에서 계산된 $\hat{g}(\theta_{normal};t)$을 이용하여 식 (17)을 통해 중심선과 관리 상한과 하한을 계산한다. 하지만 주기신호를 관리할 때 기존의 관리도와는 다르게 매 시점에 대한 Individual 관리도를 구축하여 종합적인 관리도의 형태를 사용하므로 $C.V.$ 및 ARL 계산이 어렵다. 따라서 본 연구에서 다루는 관리도는 $C.V.$값을 고정시키지 않고, 실제 신호에 대해 $C.V.$값의 변화를 통해 관리도를 구축하였다.

$$CL_t = \hat{f}(\theta_{normal};t)$$
$$UCL_t = \hat{f}(\theta_{normal};t) + C.V. \times \hat{g}(\theta_{normal};t)$$
$$LCL_t = \hat{f}(\theta_{normal};t) - C.V. \times \hat{g}(\theta_{normal};t)$$
$$(17)$$

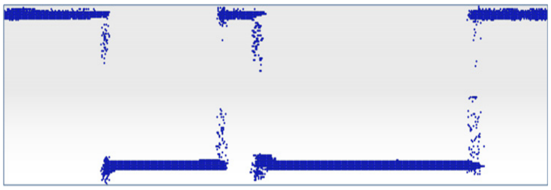
식 (17)를 사용하여 계산한 모든 시점에 대한 중심선과 관리 상한선과 하한선을 연결하면 이상탐지를 위한 SVM-R 관리도가 구축되고, <그림 9>의 점선과 같이 관리한계선을 벗어나면 이상상태일 확률이 높아진다.



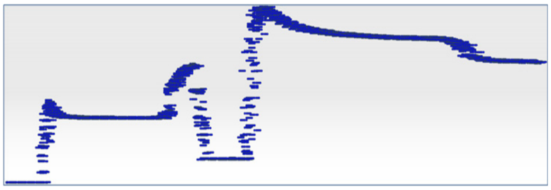
<그림 9> 제안하는 알고리즘에 대한 이상탐지의 예

3. 실험 방법 및 결과 분석

본 연구는 주기신호 이상탐지 방법에 대한 성능 개선을 위해 플라즈마(plasma) CVD(Chemical Vapor Deposition) 공정의 실제 데이터를 통하여 실험을 수행하고, 기존의 방법이 갖는 문제들을 해결하여 성능개선 효과를 확인하였다. CVD 공정이란 고순도, 고성능의 고체 물질을 생산하기 위해 사용되는 화학공정으로써 반도체나 LCD 산업에서 주로 사용된다. 성능평가 실험은 <그림 10, 11, 12>와 같이 세 가지 타입의 주기신호를 사용하였다.



<그림 10> Sensor 1의 주기신호



<그림 11> Sensor2의 주기신호

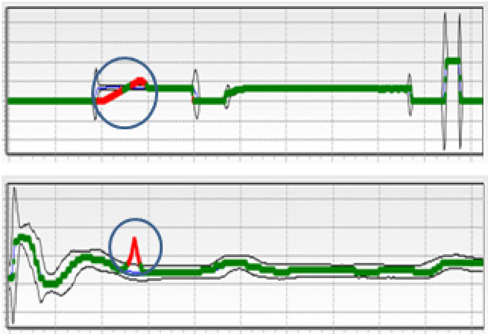
<그림 12> Sensor 3의 주기신호

세 가지 타입의 주기신호는 실제 반도체 공정이 진행되는 동안 Sensor에서 얻어진 데이터이다. Sensor 1의 주기신호는 값이 급격하게 증가하거나 감소하는 구간이 존재하는 특성을 갖고, Sensor 2는 상대적으로 값의 변화량이 적은 주기신호이다. 마지막으로 Sensor 3의 주기신호는 완만한 형태의 파형을 이룬다.

이상의 세 종류의 데이터를 바탕으로 제안된 SVM-R 관리도와 기존의 방법인 Individual 관리도를 사용한 Shewhart 관리도와의 성능을 비교하는 실험을 수행하였다.

실제 공정은 제품이 생산되며 발생하는 주기신호의 길이가 다르게 나타난다. 따라서 Shewhart 관리도를 구축 시 문제가 생기므로 신호 길이를 동일하게 맞추고 실험을 하였다. 학습데이터(training data)로는 각 Sensor의 실제 데이터 100개의 주기신호를 사용하였고, 실험(Testing)데이터로는 실제 정상데이터와 이상데이터를 7:3의 비율로 구성하여 테스트 하였다. 각 시점별로 들어오는 신호가 관리한계선을 벗어났을 경우를 이상이라고 판단한다.

흔히 반도체 공정에서 언급되는 이상상태라 함은 <그림 13 > 그래프의 원안에 표시된 부분처럼 급변하지 않고 서서히 변하는 경우나 하단 그래프의 붉은색 부분처럼 정상패턴에서 벗어난 이상신호가 발생하는 경우 등이 있다. 이는 반도체 공정에서 필요한 온도, 압력, 전압 등과 같은 값의 변동을 의미한다.



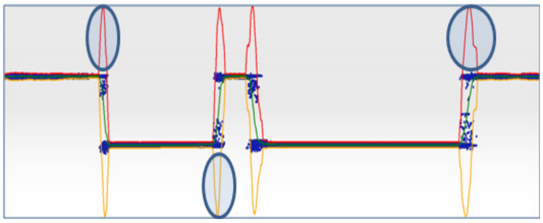
<그림 13> 주기신호의 이상상태

<그림 14, 15, 16>은 실험을 통하여 실제로 구축된 RBF 커널을 사용한 SVM-R 관리도와 Shewhart 관리도의 모습이다. 구축된 각각의 관리도를 보면 Shewhart 관리도는 각 그림에서 타원으로 표시된 부분이 분산증가로 인해 관리한계선의 증폭문제를 보인다. 따라서 상대적으로 이상탐지율이 떨어질 것으로 예측된다. 그러나 SVM-R 관리도는 Shewhart 관리도의 타원 부분 문제가 해결되었다. 이것은 SVM-R 관리도의 성능이 Shewhart 관리도에 비해 좋을 것으로 예측된다. 각 Sensor별로 구축된 관리도와 실험에 대한 결과를 살펴보면 다음과 같다.

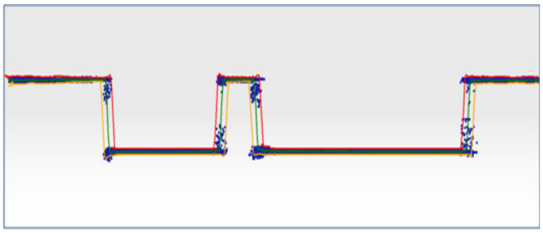
·Sensor 1 데이터

<그림 14>과 같이 Sensor 1의 데이터는 급격하게 주기신호 값이 증가하거나 감소하는 구간이 존재한다. <그림 14 - (a)>의 Shewhart 관리도는 타원으로 표시된 부분에서 상하로 값이 크게 변하여 관리한계선이 증

폭된 모습을 보인다. 반면 <그림 14 - (b)>의 SVM-R 관리도는 상하로 관리한계선이 증폭 되는 문제를 해결한다. 값이 일정한 구간은 두 개의 관리도가 비슷한 모습을 보인다.



(a) Shewhart



(b) SVM-R
<그림 14> Sensor 1에 대한 관리도

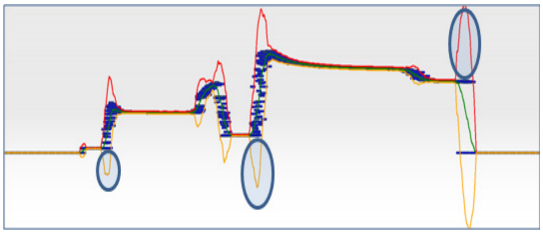
·Sensor 2 데이터

<그림 15>와 같이 Sensor 2 데이터는 Sensor 1에 비해 상대적으로 값의 변화의 형태가 다양하게 나타나며, 변화 빈도도 잦은 경우이다. 결과적으로 Sensor 2의 관리도는 Sensor 1과 비슷한 양상을 보인다. <그림 15 - (a)>의 Shewhart 관리도는 타원으로 표시된 부분처럼 상하로 값이 크게 변하는 구간에서 관리한계선이 증폭된다. 하지만 <그림 15 - (b)>의 SVM-R 관리도는 상하로 관리한계선이 증폭 되는 현상이 Shewhart 관리도에 비해 적게 나타난다. 값의 변화가 없는 구간은 두 개의 관리도가 유사한 모습을 보인다. 하지만 Sensor 1과 Sensor 2의 주기신호에서 값의 변화가 있는 구간은 관리한계선의 차이를 보인다. 이는 SVM-R 기법이 그 예측성능 비선형 주기신호의 형태나 특성에 영향을 받기 때문이다.

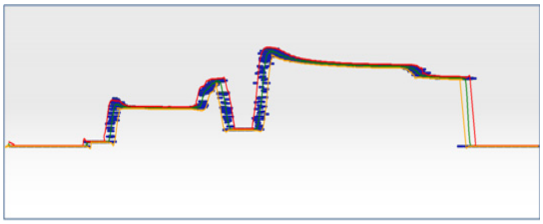
·Sensor 3 데이터

<그림 16>와 같이 Sensor 3은 Sensor 1, 2와 달리 초기에 값이 급변하는 구간을 제외하면 전반적으로 완만한 형태를 보인다. <그림 16 - (a)>의 Shewhart 관리도와 <그림 16 - (b)>의 SVM-R 관리도는 변화가 일정한 구간에서 관리한계선의 폭은 비슷하다. 하지만 Shewhart 관리도는 초기 급격한 변화 구간인 타원 부분에서 급격한 분산증가로 인해 관리한계선의 폭이 넓어지는 것을 볼 수 있다. 따라서 급격한 변화 구간의 관리한계선 폭이 Shewhart 관리도에 비해 SVM-R 관리도의 한계선이 효과적으로 설정되었다고 볼 수 있다.

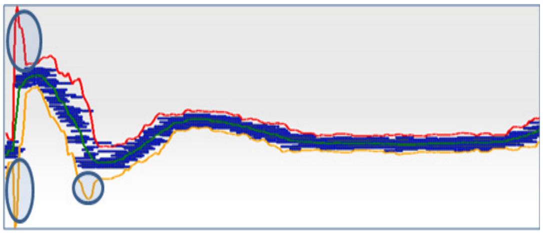
Sensor 3의 관리도는 Sensor 1과 Sensor 2에 비해 중심선과 관리한계선이 정확하게 구축되었다. 이는 평균에 분산에 영향을 주는 데이터들이 일정범위에서 고르게 존재하였기 때문이다.



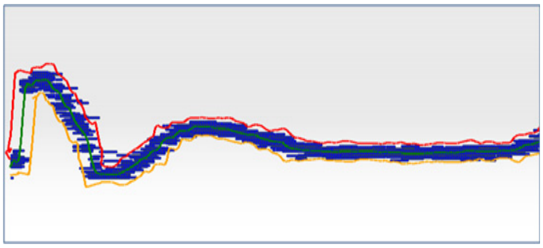
(a) Shewhart



(b) SVM-R
<그림 15> Sensor 2에 대한 관리도



(a) Shewhart



(b) SVM-R
<그림 16> Sensor 3에 대한 관리도

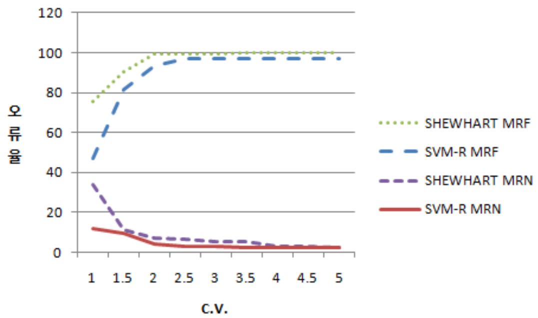
각 관리도의 탐지 성능을 확인하기 위해 사용한 100

개의 주기신호 테스트 데이터는 반도체 플라즈마 공정에서 나오는 실제 데이터로써 30개의 이상주기신호, 70개의 정상주기신호로 구성되었다. 그리고 데이터를 기존의 Shewhart 관리도와 본 연구에서 제안하는 SVM-R 관리도에 적용하였다. 일반적으로 관리도의 성능 평가를 위해 사용되는 척도는 일종오류(Type I)와 이종오류(Type II) 혹은 ARL(Average Run Length)이다. 하지만 본 연구에서 다루는 주기신호는 단 한 개의 점이 한계선을 벗어났다고 하여 해당 주기신호를 이상이라고 판단할 근거가 부족하기 때문에 기존의 관리도에 대한 성능평가 척도로써 사용되는 일종오류와 이종오류에 상응하는 MRN(Misclassification Rate for Normal signal)과 MRF(Misclassification Rate for Fault signal)를 사용하고 식으로 나타내면 다음과 같다.

$$MRN = n(fault|signal = normal) / n(signal = normal) \times 100\% \tag{18}$$

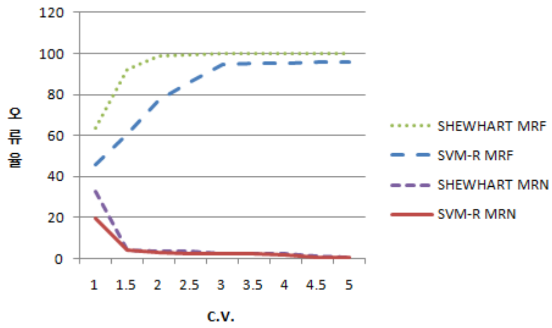
$$MRF = n(normal|signal = fault) / n(signal = fault) \times 100\% \tag{19}$$

식(18)의 MRN은 정상으로 판명된 주기신호 중에서 비정상으로 판단된 시점 데이터의 개수의 비율이고 식(19)의 MRF는 이상으로 판명된 주기신호 중에서 정상으로 판단된 시점 데이터의 개수의 비율이다. <그림 17>는 Sensor 1의 주기신호를 $C.V.$ 값을 다양하게 하여 기존의 관리도와 제안한 관리도에 적용했을 때 계산된 오류율을 보여준다. 전체적으로 $C.V.$값에 관계없이 SVM-R 관리도의 MRF과 MRN은 기존의 관리도 보다 낮게 나타났다. 하지만 $C.V.$값이 2보다 큰 구간에서는 오류율의 변화가 적었다.



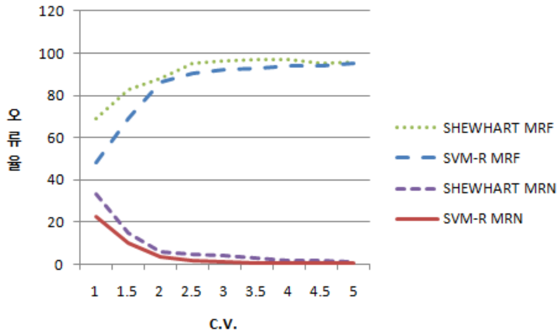
<그림 17>Sensor 1 주기신호의 탐지오류율

<그림 18>은 Sensor 2의 주기신호를 적용한 결과를 나타낸다. $C.V.$가 1.5보다 큰 구간에서 MRN의 크기가 비슷한 것을 제외하고는 전반적으로 SVM-R 관리도의 MRN과 MRF가 적게 나타났다.



<그림 18>Sensor 2 주기신호의 탐지오류율

<그림 19>의 Sensor 3의 경우는 $C.V.$값에 관계없이 MRF과 MRN의 값은 제안한 SVM-R 관리도에서 더 낮게 나타났다. 종합적으로 봤을 때 $C.V.$값에 관계없이 제안하는 관리도의 오류율이 적게 나타나는 경향을 볼 수 있고, $C.V.$값이 작은 경우 오류율이 특히 적은 것을 알 수 있다. MRF의 수치가 MRN에 비해 큰 값을 갖는 이유는 전체 주기신호를 구성하는 점들에 비해 이상으로 판단되는 점들의 개수가 극히 작기 때문이다. 따라서 이상임에도 불구하고 정상으로 판단되는 점의 개수가 급격히 커지기 때문에 MRF는 큰 값을 갖는다.



<그림 19>Sensor 3 주기신호의 탐지오류율

4. 결론 및 추후연구

본 연구는 SVM-R을 이용한 관리도를 구축하여 이상탐지 방법을 제안하였다. 주기신호의 길이가 각기 다른 경우에는 Shewhart 관리도를 구축하는데 어려움이 있기 때문에 SVM-R을 사용하였다. 또한 신호의 변화가 큰 구간에서의 분산증가 현상으로 인한 문제를 해결하기 위해 비선형 주기신호에 대한 분산을 추정하는 함수를 정의하여 관리도를 구축하였다.

실제 공정데이터를 사용하여 기존 관리도와의 성능

평가를 수행하였으며, 본 연구에서 제안한 알고리즘의 뛰어난 이상탐지능력을 확인하였다.

따라서 본 연구가 실용화될 경우 불량률을 줄이고 공정의 효율을 보다 높일 수 있을 것으로 기대한다.

하지만 주기신호에 대한 관리도에서 중심선과 관리한계선을 보다 정확하게 구축하기 위해 SVM-R의 파라미터를 적절히 조절하여 최적의 중심선을 찾을 필요가 있다. 또한 주기신호의 급격한 변화 구간에서는 분산을 보다 정확히 추정할 수 있는 방법에 대한 연구가 필요하다. 또한 여러 Sensor를 동시에 관리할 수 있는 다변량 관리도에 대한 연구가 필요하다.

감사의 글

·이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2010-0075383).

·본 과제는 정보통신산업진흥원의 SW공학 요소기술 개발과 전문인력 양성사업의 결과물임을 밝힙니다.

·이 논문은 2010년도 2단계 두뇌한국(BK)21사업에 의하여 지원되었음.

참고문헌

[1] 박찬규(2006), "Support Vector Regression을 이용한 소프트웨어 개발비 예측", 「한국경영과학회지」, 23권, 2호, pp. 75-91.

[2] Burges, C. J. C.(1998), "A tutorial on support vector machines for pattern recongnition", *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121-167.

[3] Montgomery, D. C.(2001), *Introduction to Statistical Quality Control*, *5th Edition*, Hohn Wiley & Sons, NewYork, NY.

[4] Nello, C. and John, S.(2000), *An Introduction to Support Vector Machine*, Cambridge.

[5] Shi, J. and Jin, J.(2000), "Diagnostic feature extraction from stamping tonnage signals based on design of experiments", *Journal of Manufacturing Science and Engineering*, Vol. 122, No. 22, pp. 360-369.

[6] Smola, A., Scholkopf. B.(2004), "A tutorial on support vector regression", *Statistics and Computing*, Vol. 14, pp. 199-222.

[7] Vapnik, V.(1995), *The Nature of Statistical Learning Therory*, Springer-Verlag.

[8] Vapnik, V.(1998), *The Statistical Learning Theory*, John Wiley & Sons, Inc.

[9] http://www.public.itrs.ne

2010년 6월 3일 접수, 2010년 9월 3일 1차 수정, 2010년 9월 14일 2차 수정, 2010년 9월 15일 채택