생성적 적대 신경망을 이용한 반도체 제조공정 데이터의 결측치 추정 및 공정 이상 진단 프레임워크

Fault Detect and Classification Framework for Semiconductor Manufacturing Processes using Missing Data Estimation and Generative Adversary Network

저자 (Authors)
김희수, 이현수
Heesoo Kim, Hyunsoo Lee

출처 (Source)
한국지능시스템학회 논문지 28(4), 2018.8, 393-400(8 pages)
Journal of Korean Institute of Intelligent Systems 28(4), 2018.8, 393-400(8 pages)

발행처 (Publisher)
한국지능시스템학회
Korean Institute of Intelligent Systems

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07520681

APA Style
김희수, 이현수 (2018). 생성적 적대 신경망을 이용한 반도체 제조공정 데이터의 결측치 추정 및 공정 이상 진단 프레임워크. 한국지능시스템학회 논문지, 28(4), 393-400

Journal of Korean Institute of Intelligent Systems
Vol. 28, No. 4, August 2018, pp. 393–400
http://dx.doi.org/10.5391/JKIIS.2018.28.4.393

ISSN(Print) 1976–9172
ISSN(Online) 2288–2324

JKIIS

생성적 적대 신경망을 이용한 반도체 제조공정 데이터의 결측치 추정 및 공정 이상 진단 프레임웍

Fault Detect and Classification Framework for Semiconductor Manufacturing Processes using Missing Data Estimation and Generative Adversary Network

김희수 · 이현수†
Heesoo Kim and Hyunsoo Lee†

국립 금오공과대학교 산업공학부
School of Industrial Engineering, Kumoh National Institute of Technology

요 약

많은 현대 제조 공정에서 선진화된 센서를 활용한 공정 이상 유무 진단 프레임웍이 활용되고 있다. 이때, 일부 장비의 설비 보존, 정비, 고장 등으로 인하여 측정되는 센서 데이터의 결측치가 존재하게 되는 이슈가 발생한다. 또한, 제조 공정의 고도화는 불합격 표본 데이터의 부족을 초래하여, 학습 데이터의 불균형을 초래하게 된다. 이러한 문제점을 개선하기 위하여 본 연구에서는 데이터의 분포에 기반 한 샘플링을 통하여 결측치를 추정하고, 이를 다시 생성적 적대 신경망을 이용하여 보정하는 프레임웍을 제안한다. 또한, 제안된 생성적 적대 신경망을 통하여 불균형한 표본 데이터를 생성데이터로 보완하여 균형된 데이터의 학습이 이루어지게 한다. 이를 위하여 실제 반도체 제조공정 데이터를 통해 제안된 프레임웍을 테스트하고, 기존 연구들과의 비교를 통하여 그 우수성을 증명한다.

키워드 : 반도체 제조 공정 빅데이터, 결측치 추정 및 보정, 생성적 적대 신경망, 데이터 마이닝, 공정이상 진단

Abstract

Contemporary manufacturing processes have introduced several fault detection and classification systems using advanced sensor technologies. However, the maintenances and breakages of the manufacturing facilities make it possible to generate the missing values in the process data. In addition, the advanced process technologies give result to the shortage of the fault data and then, this issue causes imbalances of the data. In order to overcome these issues, the paper estimates the missing values using the data distribution based Monte Carlo based sampling method and then these data are interpolated using the proposed generative adversary network. The used network produces the more virtual fault data for balancing the input data. In order to show the proposed framework, the real manufacturing process data are tested and the applied framework is compared with the existing methods.

Key Words : Semiconductor Manufacturing Process Big Data, Missing Value Estimation and Interpolation, Generative Adversary Network, Data Mining, Fault Detection and Classification

Received: Jul. 23, 2018
Revised: Aug. 10, 2018
Accepted: Aug. 13, 2018
†Corresponding authors
hsl@kumoh.ac.kr

본 논문은 교육 과학 기술부의 한국연구재단(NRF) 기금을 통한 기초과학 연구 프로그램에서 지원하여 연구하였음 (grant number: NRF–2018R1D1A3B07047113).
This research was supported by The Basic Science Research Program through the National Research Foundation of Korea (NRF) fund by the Ministry of Education, S. Korea (grant number: NRF-2018R1D1A3B07047113)."

1. 서 론

제조 공정에서 이상감지 및 분류를 위해 기계학습이나 딥러닝을 적용하여 문제를 해결하고자 하는 연구들이 활발히 이어지고 있다. 특히 10나노 수준의 반도체를 생산하기 위한 제조공정에서는 불량발생의 최소화와 생산 효율 증대를 위해 각 공정에 관련 센서를 부착하여 상태를 모니터링하는 것이 표준 프로세스로 자리잡아 왔다. 이러한 센서 측정치를 통하여 이상감지 및 분류를 통한 빠른 조치[1]나 예방적 조치를 하는 것은 현대 제조공정의 중요한 특징으로 자리잡아 왔다. 이때 공정 이상 측정에 많은 기계학습 방법들이 많이 적용하는데, 이때 데이터가 충분히 많고, 클래스 별로 잘 분포되어 있을 때 좋은 성능을 낸다. 하지만 실제 제조공정에서 얻어진 데이터는 이러한 데이터가 불균형적으로 분포되어 있으며, 이러한 데이터로 학습한 모델은 특정 클래스에 편향되어 학습되고 이는 최종적인 분류성능에 영향을 미치게 된다[1,2,3]. 또한 실시간의 센서기반 공정에서 얻어진

데이터는 각 구성항목에서 다수의 결측치 [4]를 포함하기도 한다.

이러한 데이터의 불균형 문제를 해결하기 위해 소수 클래스의 오버샘플링 기법 (Synthetic Minority Over Sampling Technique; SMOTE)이 사용되기도 하며, 결측치 대체 방법으로는 mean/mode, 회귀모델 [4,5]등이 사용되기도 한다. 오버샘플링 기법인 SMOTE는 고차원데이터에서 효과적이지 못한 단점을 지니고 있으며[6], 다수의 결측치 대체 방법으로의 회귀모델은 과적합의 문제점을 가지고 있다. 또한, 평균/최빈값의 대입은 결측치 대체의 빠르고 간단한 접근 방법[5,7]이지만, 정확도 측면에서 문제의 여지를 가진다.

본 연구에서는 반도체 생산라인으로부터 획득된 SECOM [8] 데이터 셋을 사용하여 생성적 적대 신경망 (Generative Adversarial Networks; GAN)을 통해 소수 클래스의 오버샘플링을 수행하고, 이를 이상감지 및 분류를 위한 훈련데이터로 포함시킨 후 합격 또는 불합격의 분류를 하는 프레임웍을 제안한다.

이를 위하여 다음 장에서는 본 연구의 핵심이 되는 결측치 처리에 관한 연구와 GAN을 통한 오버샘플링에 관한 연구를 살펴보고, SECOM데이터를 활용한 여러 분류모델을 비교 분석한다. 3장에서는 결측치를 보정하면서 오버샘플링 및 분류 모델을 제시하며, 4장에서 SECOM 데이터를 사용한 시뮬레이션을 통하여 제안된 프레임웍의 우수성을 증명한다.

2. 배경 및 문헌연구

2.1 결측치 처리 연구

나노 공정의 반도체 공정은 기존의 반도체 공정에 비하여 보다 높은 복잡도를 지니며, 특히 수율 관리부문에서 대부분의 이슈가 집중된다. 따라서, 높은 수준의 수율 획득을 위해 선제적인 공정에서의 이상감지 및 분류 [9]가 중요하다. 이를 위해 대부분의 공정은 센서로부터 공정변수를 일정 시간간격으로 측정하며 모니터링[1]한다. 하지만, 센서의 고장 및 설비 셋업, 정비주기와 함께 측정되는 공정데이터의 부분에서 결측치가 존재하게 된다.

따라서, 이를 고려한 데이터의 분석은 결측치를 고려한 처리가 반드시 선제되어야 한다. 결측된 값의 처리는 대부분 대체 값을 채우거나 제거함으로써 처리된다. 불완전한 속성을 제거하는 방식은 결측된 수가 상대적으로 적고 영향을 거의 미치지 않을 때 사용하며, 대체적으로는 결측치 값의 추정 및 대체방법[10]을 사용한다. 일반적인 결측치 보정방법으로는 평균/중앙값 대체가 있으며, 이외에도 k-Nearest Neighbor (K-NN), 인공신경망(Neural Network), 회귀모델(Regression) 등을 통한 대체방법[11]등도 있다. Yuan [12]은 단일 추정값으로의 대체가 아닌 불확실성을 가진 대체 방법인 Multiple Imputation을 제안하였고, Kerdprasop K.와 Kerdprasop N. [13]은 데이터마이닝 기법 중 하나인 Problem Rule Induction Method(PRIM)에서 누락된 값의 체계적인 처리와 프로세스 개선을 위한 Missing-value PRIM 을 제안하였다.

본 연구에서는 GAN 구조에서 학습 및 역전파 알고리즘을 통해 결측치를 처리하는 방법을 제안하고 그 유효성을 증명한다.

2.2 오버샘플링을 통한 이상감지 및 분류

일반적인 기계학습에서는 각 클래스 별 샘플수가 대략적으로 비슷[14]하다고 가정하고, 그 편향성을 없애고자 한다. 하지만, 실제 제조 현장에서 얻어진 데이터는 일부의 클래스에 편향되어

표 1. 이상감지 및 분류 프레임웍과 방법 비교

Table 1. Comparisons of Fault Detection and Classification Frameworks and Methods

	Data Imbalance	Feature Selection	Missing value Imputation	Classification
K.Kerdprasop, N.Kerdprasop [17]	Duplicate fail case	PCA, Chi-Square, Remove features(containing missing more than 55%)	X	K-NN, Logistic Regression, Naïve Bayes, Decision Tree
D.Moldovan, T.Cioara [18]	SMOTE	PCA, Bourta, MARS, Remove feature (containing missing more than 55%)	Mean, nearest neighbor	Logistic Regression, Random Forest, Gradient Boosted Trees
S.Munirathinam , B.Ramadoss [19]	X	CA, PCA , Remove feature (containing missing more than 55%)	X	ANN, SVM, Naïve Bayes, Decision Tree, K-NN, Logistic Regression
G.Douzas, F.Bacao [20]	cGAN(image)	X	X	Logistic Regression, SVM, K-NN, Decision Tree, GBM
Q. Peter He, Jin Wang [21]	X	PCA,MPCA	X	FD-kNN
Son, Ko, and Kim [1]	SMOTE	Global Feature, Local Feature	X	Decision Tree
Lee and Hwang [16]	GAN	EMD(Empirical mode decomposition)	X	Binary Classification
Nam and Kim [15]	SMOTE	Selection by experience	X	Decision Tree, SVM, Logistic Regression, ANN, Random Forest

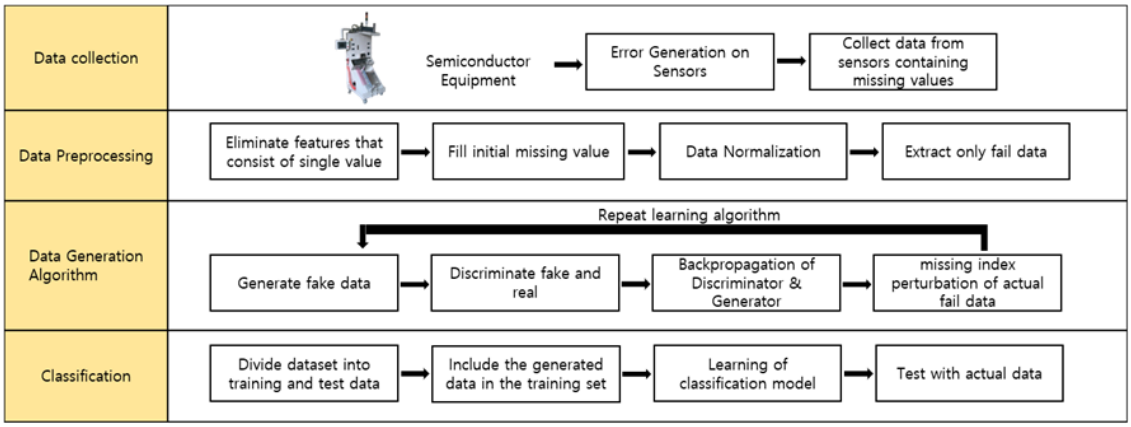


그림 1. 결측치 보정 및 오버샘플링을 통한 이상감지 및 분류 프레임워크

Fig. 1. Fault Detection and Classification Framwork using the proposed method

있는 경우가 많다. 이처럼 불균형 데이터의 사용은 학습 및 그 결과 적용의 신뢰도를 떨어뜨리는 요인이 된다. 특히, 반도체 공정에서 제품의 수율이 안정화되면서 고수율 웨이퍼와 저수율 웨이퍼의 비율이 심각한 불균형 상태[15]가 된다. 그러므로 분류모델을 구축하기 위해서는 데이터 불균형 문제를 해결하는 것이 가장 중요한 선결조건이 된다. Son, Ko와 Kim [1]은 오류검출 및 분류(Fault Detection and Classification; FDC)을 위해 공정결과와 연관성이 높은 데이터 스트림의 구조적인 특징들을 추출하여 의사결정 트리의 입력으로 사용하여 분류하는 모델을 제안하였다. 표 1은 기존 연구들의 이상감지 및 분류를 위한 알고리즘 비교 및 프레임웍을 보여준다. 표 1에서 볼 수 있듯 대부분의 프레임웍은 결측 값을 고려하지 않고 특징선택을 통한 분류 프레임웍을 기술한다. Lee와 Hwang[16]은 GAN을 사용하여 데이터를 오버샘플링 한 후 분류모델에 적용하였지만 여전히 반도체 공정에서 흔히 발생하는 결측치에 대한 고려 없이 가상데이터를 생성하였다.

대부분의 기존 문헌연구들에서는 이상감지 및 분류 문제에서 데이터 불균형을 해결하기 위해 즉, 불량이나 불합격의 소수데이터를 해결하기 위해서 오버샘플링 기법을 적용하고 특징을 추출하여 불량의 분류모델을 제안하였다. 하지만 대부분의 결측치는 그 처리 대상에서 제외하거나 그래도 남아있는 결측치는 평균/중앙값 대체하는 방법을 사용하였다. 장비나 센서의 오류로부터 발생하는 결측치를 고려하지 않은 오버샘플링을 통해 분류모델을 구축하는 것은 잠재적으로 가치 있는 변수를 탈락시키는 한계점을 지닌다.

본 연구에서는 신경망으로 구성된 생성모델인 GAN을 사용하고, 학습알고리즘에서 결측치에 대한 처리과정을 접목하여 미세한 결측치의 보정을 통한 오버샘플링을 기반의 제품 이상을 분류하는 효율적인 방법을 모색한다.

3. GAN을 활용한 결측치 보정 및 생성을 통한 제조공정 이상탐지 프레임웍

제조공정에서 수많은 센서로부터 얻어진 측정값은 많은 결측치를 포함하거나 클래스마다 샘플 수의 불균형과 같은 문제[2]를 가지고 있다. 이러한 샘플 수의 불균형이나 누락된 값을 포함한 자료는 데이터의 정확한 분석을 어렵게 만드는 요인[13,18]이 된다. 분류문제에서 결측치를 포함한 데이터를 적용할 때, 일반적으로 데이터 전처리 과정에서 결측치가 많은 특징을 제거하거나, 결측치를 대체하여 완전한 데이터로 만든 후 생성모델[5]에 적용한다.

본 연구에서는 불완전한 자료 처리를 위해 모든 결측치에 대한 대체 및 미세한 보정을 실시하고, 생성 모델인 GAN을 사용하여 훈련 데이터와 유사한 가상 불합격 데이터를 생성한다. 또한, 이렇게 생성된 데이터를 합격과 불합격의 분류를 위한 훈련 데이터로 사용하여 모델 학습을 하는 프레임웍을 제안한다. 그림 1은 제안하는 모델 생성 및 분류의 전체적인 과정을 보여준다. 다음의 하위 절은 각 단계별 세부사항을 나타낸다.

3.1 GAN을 이용한 결측치의 보정을 동반한 소수 클래스의 오버샘플링

1) 데이터 전처리

먼저 데이터셋을 모델에 적용하기 전에 전처리 과정을 수행한다. 전처리 과정으로 결측치의 대체 및 데이터의 정규화를 한다. 이를 수행하기 위하여, 데이터 분포에 영향을 미치지 않는 상수 값 형태의 데이터 속성값을 제거한다. 기존 연구들에서는 추가적으로 결측치의 개수가 샘플 수의 45~60%를 넘으면 그 특징에서는 유의미한 정보를 얻을 수 없다 판단하고 그 특징들을 제거[17,19,22]하지만, 본 연구에서는 잠재적으로 가치 있을 수 있는 특징들을 모두 고려하기 위해 추가의 특징제거과정은 제외한다.

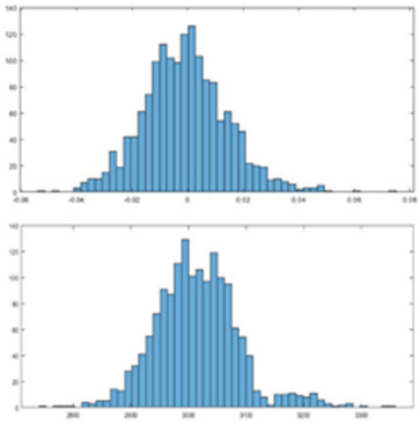


그림 2. 특징값들의 데이터 분포

Fig. 2. Data fitting of feature vectors

다음으로 'NaN'이라고 표기되어 있는 결측치를 초기값으로 대체하는 과정을 수행한다. 각 데이터의 속성값들을 그림 2와 같이 각 데이터의 특성에 맞는 분포로 모델링하고, 이를 통해 난수를 발생시켜 초기값을 대체한다. 식 (1)은 특성값이 정규분포를 따를 때 초기값을 대체하는 과정을 기술한다. 식(1)에서 i는 특징, j는 각 특징의 속성값 그리고 k는 특징벡터에서의 결측을 의미한다.

$$\begin{aligned} x_i &= \{x_{i,1}, x_{i,1}, \cdots, x_{i,k}, x_{i,k+1}, \cdots, x_{i,j}\} \\ x_{i,1}^{*} &\leftarrow x_{i,1} \sim N(\mu_i, \sigma_i) \\ x_i &= \{x_{i,1}, x_{i,1}, \cdots, x_{i,k}^{*}, x_{i,k+1}, \cdots, x_{i,j}\} \\ (\mu_i^{*}, \sigma_i^{*}) &\leftarrow (\mu_i, \sigma_i) \\ x_{i,1}^{*} &\leftarrow x_{i,1} \sim N(\mu_i, \sigma_i) \\ x_i &= \{x_{i,1}, x_{i,1}, \cdots, x_{i,k}^{*}, x_{i,k+1}^{*}, \cdots, x_{i,j}\} \end{aligned} \tag{1}$$

그 다음 얻어진 측정 값들이 다양한 범위를 가지므로, 데이터의 처리속도 및 학습 수렴속도를 위하여 [-1,1]의 범위로 값의 정규화 [19,23]를 수행한다(식 (2)).

$$\hat{x_i} = \frac{x_i - \mu_i}{\sigma_i} \tag{2}$$

2) GAN을 활용한 결측치 보정 및 데이터 생성

본 절에서는 데이터를 생성하기 위한 알고리즘은GAN을 사용한다. 생성데이터의 클래스는 상대적으로 데이터의 수가 적은 불량 데이터를 대상으로 한다.

GAN은 생성자 G(Generator)와 구별자 D(Discriminator)의 대립구조의 학습을 통해 훈련데이터와 유사한 자료를 만들어내는 생성모델이다. 이때 D와 G는 미분 가능한 함수로, 각각이 $\theta^{(D)}, \theta^{(G)}$를 파라미터로 갖는 멀티레이어 퍼셉트론이다. 멀티레이어 퍼셉트론 구조에서 가중치(weight) w_i , 각 레이어의 입력 y_{i-1} 그리고 바이어스 b_i 가 있을 때, 각 은닉층 유닛 값은 식 (3), (4)을 따른다.

$$v_i = (w_i \cdot y_{i-1}) + b_i \tag{3}$$

$$y_i = \varphi(v_i) \tag{4}$$

이때 $\varphi(\cdot)$은 사용자가 설정한 활성화 함수이며, 첨자 i는 i번째 레이어를 의미한다. 이러한 구조를 바탕으로 D와 G는 각각 식(5)의 최대화하면서 동시에 두 번째 항을 $f_D(x^g)$ 최소화 하도록 학습된다.

$$\begin{aligned} \min_G \max_D V(D,G) &= E_{x \sim p_{data}(x)}[\log D(x)] \\ &\quad + E_{z \sim p_{(z)}}[\log(1 - D(G(z)))] \end{aligned}$$

$$\min_G \max_D V(D,G) = f(x^r) + f(x^g) \tag{5}$$

$$\begin{aligned} Where: f_D(x^r) &= E_{x \sim p_{data}(x)}[\log D(x)] \\ f_G(x^g) &= E_{z \sim p_{(z)}}[\log(1 - D(G(z)))] \end{aligned}$$

식(5)에서 $p_{data}(x)$는 실제 데이터 분포를 의미하며 $p(z)$는 생성자의 입력 잡음 분포를 의미한다. 최적화는 모멘텀 방식 [20]을 적용하여 D와 G는 각각 1과 0이 되도록 학습한다. 그림 3은 역전파 알고리즘을 사용하여 각 파라미터의 그레디언트를 산출 (식 (6-9)) 하고, 이를 통해 학습하는 과정을 도식화하여 보여준다.

식(6)~(8)은 구별자의 역전파, 식(9)는 생성자의 역전파 계산과정을 의미한다.

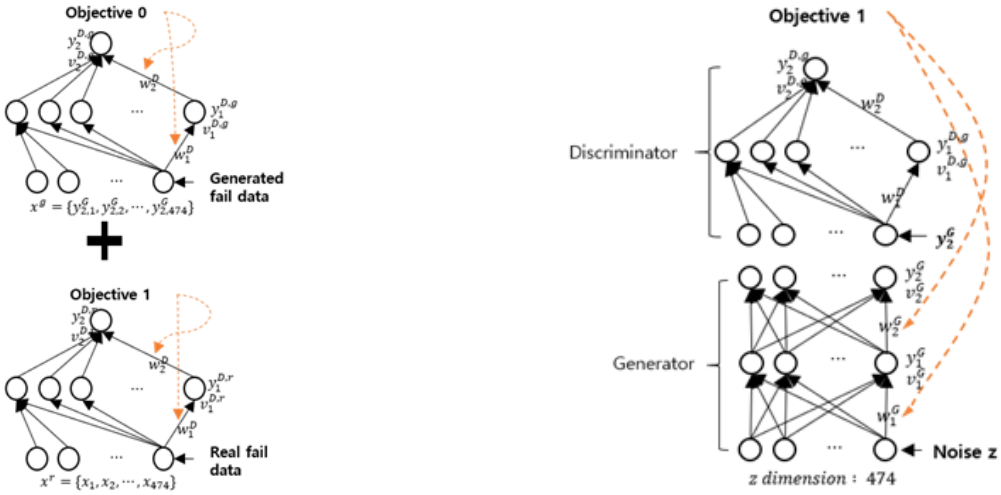
$$\frac{\partial V}{\partial \theta^{(D)}} = \frac{\partial f(x^r)}{\partial \theta^{(D)}} + \frac{\partial f(x^g)}{\partial \theta^{(D)}} \tag{6}$$

$$\begin{aligned} \frac{\partial f_D(x^r)}{\partial \theta_k^{(D)}} &= \frac{\partial f_D(x^r)}{\partial y_i^{D,r}} \cdot \frac{\partial y_i^{D,r}}{\partial v_i^{D,r}} \cdot \frac{\partial v_i^{D,r}}{\partial y_{i-1}^{D,r}} \cdot \cdots \cdot \frac{\partial v_k^{D,r}}{\partial \theta_k^{(D)}} \\ \frac{\partial f_D(x^r)}{\partial w_1^{(D)}} &= \frac{1}{D(x)} \cdot \varphi'(v_2^{D,r}) \cdot w_2^{D} \cdot \varphi'(v_1^{D,r}) \cdot x^r \end{aligned} \tag{7}$$

$$\begin{aligned} \frac{\partial f_D(x^g)}{\partial \theta_k^{(D)}} &= \frac{\partial f_D(x^g)}{\partial y_i^{D,g}} \cdot \frac{\partial y_i^{D,g}}{\partial v_i^{D,g}} \cdot \frac{\partial v_i^{D,g}}{\partial y_{i-1}^{D,g}} \cdot \cdots \cdot \frac{\partial v_k^{D,g}}{\partial \theta_k^{(D)}} \\ \frac{\partial f_D(x^g)}{\partial w_1^{(D)}} &= \frac{1}{1 - D(G(z))} \cdot \varphi'(v_2^{D,g}) \cdot w_2^{D} \cdot \varphi'(v_1^{D,g}) \cdot x^g \end{aligned} \tag{8}$$

$$\begin{aligned} \frac{\partial V}{\partial \theta_k^{(G)}} &= \frac{\partial f(x^g)}{\partial y_i^{D}} \cdot \frac{\partial y_i^{D}}{\partial v_i^{D}} \cdot \cdots \cdot \frac{\partial v_1^{D}}{\partial x^g} \cdot \frac{\partial x^g (= y_i^{G})}{\partial v_i^{G}} \\ &\quad \cdot \frac{\partial v_i^{G}}{\partial y_{i-1}^{G}} \cdot \cdots \cdot \frac{\partial v_k^{G}}{\partial \theta_k^{G}} \\ \frac{\partial V}{\partial w_1^{(G)}} &= \frac{1}{D(G(z))} \cdot \varphi'(v_2^{D,g}) \cdot w_2^{D} \cdot \varphi'(v_1^{D,g}) \\ &\quad \cdot w_1^{D} \cdot \varphi'(v_2^{G}) \cdot w_2^{G} \cdot \varphi'(v_1^{G}) \cdot z \end{aligned} \tag{9}$$

식 (6)에서 과 는 각각 실제 훈련데이터와 생성된 가상 데이터를 의미한다. 구별자 D의 학습은 비용함수를 D의 파라미터 에 대해 각각 편미분하여 산출되고, G의 학습은 D로부터 산출된 오류를 전달받아 그 그래디언트를 구한다. 한 반복실험 내에서 D와 G의 파라미터 업데이트가 되고 나면 다시 결측치에 대한 보정을 한다.



(a)구별자의 역전파 흐름
(a) Backpropagation of discriminator

(b) 생성자의 역전파 흐름
(b) Backpropagation of generator

그림 3. 생성모델 GAN의 역전파 알고리즘 흐름
Fig. 3. Backpropagation algorithm flow of the used GAN

결측지의 재보정에 관한 모델링은 다음 식 (10-11)과 같이 한다.

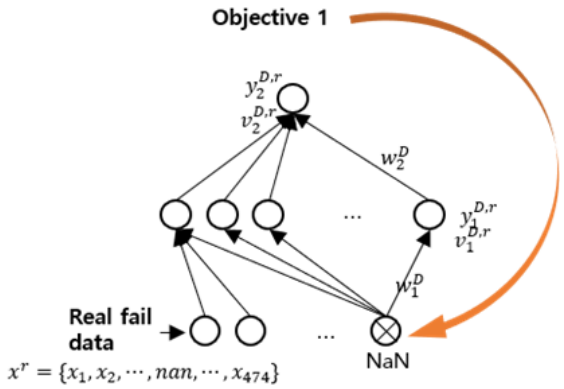


그림 4. 결측 값의 미세조정 과정
Fig. 4. Pertubation process for interpolating missing values

$$x^r(m) = x^r(m) + \epsilon \cdot \Delta x^r(m) \tag{10}$$

$$\frac{\partial f_D(x^r)}{\partial x^r} = \frac{\partial f_D(x^r)}{\partial y_i^{D,r}} \cdot \frac{\partial y_i^{D,r}}{\partial v_i^{D,r}} \cdot \frac{\partial v_i^{D,r}}{\partial y_{i-1}^{D,r}} \cdot \cdots \cdot \frac{\partial v_1^{D,r}}{\partial x^r} = \frac{1}{D(x)} \cdot \varphi'(v_2^{D,r}) \cdot w_2^D \cdot \varphi'(v_1^{D,r}) \cdot w_1^D \tag{11}$$

식 (10)에서 m은 누락된 값의 인덱스를 의미하며, ϵ는 유의미한 영향을 주기 위한 파라미터 [24]로써 역할을 하며 실험자에 의해 결정된다. 을 결측치에 대한 그래디언트에 곱해주는 섭동항 $\Delta x^r(m)$으로 결측치의 업데이트 및 학습이 이루어진다.

훈련데이터에서 전처리 과정으로 초기값을 정규분포로부터 샘플링한 값에 보정을 거치게 되면서 이는 다음 반복실험에서 모델의 파라미터 $\theta^{(D)}$와 $\theta^{(G)}$의 학습에 간접적인 영향을 미치게 된다. 이러한 결측치의 미세 보정을 통하여 생성되는 결측치의 대체값은 실제 데이터 분포에서 발생한 데이터와 매우 유사하게 생성된다. 그림 4는 결측치의 미세조정과정을 설명한다.

다음 하위 절은 위의 방법으로 생성한 데이터를 합격/불합격 분류문제의 훈련데이터로 사용하는 것에 대한 설명을 나타낸다.

3.2 합격/불합격의 분류

머신러닝을 통해 제조데이터의 분류문제에서 중요한 이슈 중의 하나는 데이터의 불균형 문제와 결측치의 처리이다. 소수 클래스에 대해 오버샘플링과 언더샘플링의 방법이 있는데 본 연구에서는 결측치 보정을 통한 오버샘플링으로 문제 해결에 접근하였다.

생성모델 GAN을 통해 발생한 데이터가 유의미한 성능을 내는지 평가하기 위해 합격과 불합격의 분류문제에서 훈련 데이터로써 생성한 가상데이터를 분류기에서 학습하고, 테스트에서 실제 데이터만으로 분류를 함으로서, 그 모델성능을 평가한다. 이때 학습과정에서 성능의 향상을 위해 생성한 가상 불합격데이터와 실제 불합격데이터 약 20%를 포함하여 학습한다.

모델의 평가는 혼동행렬(Confusion Matrix)를 사용하여 True positive rate(TPR), $TPR = \frac{TP}{TP+FN}$ False positive rate(FPR)은 $FPR = \frac{FP}{FP+TN}$, 정밀도(Precision)은 $Precision = \frac{TP}{TP+FP}$, F-measure은 $F-measure = \frac{2TP}{2TP+FP+FN}$로 계산하기, 이를 각 성능지표로 사용한다. 여기서 TPR은 참이라고 예측된 것 중 실제

참인 것의 비율, FPR은 거짓을 참이라고 분류한 비율, 정밀도는 참이라고 예측한 것 중 실제 참인 비율을 뜻하며 마지막으로 F-measure은 정밀도와 TPR의 조화평균으로 계산된 성능지표로써 표2를 통해 산출된다.

표 2. 성능 평가를 위한 혼동행렬

Table 2. Confusion Matrix for evaluating the classification performance

	Predicted Class		
Actual Class	Value	Class=1(fail)	Class=-1(pass)
	Class=1	TP	FN
	Class=-1	FP	TN

4. 실험 및 분류 성능 결과

4.1 가상데이터로 학습한 분류기의 성능

실험에 사용한 데이터는 SECOM 데이터 셋[8]으로 반도체 제조공정으로부터 발생한 데이터의 오픈 데이터로서 많은 기존 연구에서 사용되었다. 개체 수 1567개, 생산라인 설비의 수많은 센서로부터 얻어진 특징벡터는 591개이며 각 개체의 판별은 합격과 불합격 두 가지로 나뉜다. 합격과 불합격은 각 -1/1로 나타나 있다. 본 연구의 실험에 사용하기 위해 전처리 과정을 거쳐 상수값을 가지는 특징벡터 117개를 제거하고 총 474개의 특징만을 고려하였다. 이 데이터는 합격데이터 1463개(93.3%), 불합격데이터가 104개(6.6%)로 데이터의 불균형 문제를 가지고 있으며, 이러한 불균형 문제 및 결측치를 위하여 본 연구에서 제안한 GAN을 통한 데이터 생성방법을 적용하였다.

실험을 위한 하이퍼 파라메터로 학습률(learning rate)은 0.00001, 모멘텀계수(beta)는 0.5, 구별자 D에서의 활성화 함수는 계수 2를 가진 relu함수, 미니배치(mini-batch)는 4로 설정하였다. 비교분석을 위해 같은 조건으로 결측치의 미세 보정 없이 데이터를 생성한 후 출력층의 활성화 함수 값과 목표 값의 차이를 그림 4에 표시하였다.

그림4와 같이 결측치의 미세조정을 거친 생성모델이 목표결과에 더 잘 수렴하고, 실제 목표 값에 가깝게 결과를 내는 것을 알 수 있다.

다음 하위 절에서는 명확한 성능 비교를 위해 결측치 처리를 대상으로 기존 알고리즘과 제시된 알고리즘을 비교한다.

4.2 기존 SECOM데이터를 활용한 기존 연구결과들과의 비교

본 절에서는 SECOM데이터를 사용한 기존의 분류모델과 본 논문에서 제안된 학습방법으로 분류를 수행한 결과를 각각 비교분석하고, 결측치를 통한 오버샘플링을 하였을 때 분류가 효과적으로 이루어짐을 보여준다.

비교대상은 SECOM데이터의 활용 연구에서 가장 좋은 분석결과를 보여주었던 군집 기반의 특징추출 기법인 MeanDiff 방법과 의사결정트리를 결합한 기법 [17]이다. 비교대상 방법에서는 결측치가 55%이상 차지하는 속성은 제거하였으며 남은 결측된 값에 대한 처리방법은 고려되지 않고, 특징 168개만을 선택하여 소수 그룹의 샘플 수를 복제하여 오버샘플링한 후 분류 알고리즘에 적용하였다.

이에 비해 본 연구에서는 잠재적 분류 가치가 있을 수 있는 특징 474개를 모두 고려하였다. 비교대상과 본 연구에서 제안한 모델의 성능평가를 위해 혼동행렬을 작성하고 결과를 비교하였다. 다음의 표 3의 (b)~(e)는 비교대상을 비롯한 여러 기법의 혼동행렬[17]을 보여주며 (a)는 제안한 모델의 결과를 보여준다.

제안된 모델의 합격클래스에 대한 정확도는 100%으로 분류를 잘 한다고 볼 수 있으며, 합격한 제품을 불합격이라고 분류할 확률은 0.0%으로 나타났다. 표 4의 성능지표 비교를 통해 기존 프레임웍들보다 제안된 프레임웍의 효과가 더 우수함을 보여준다.

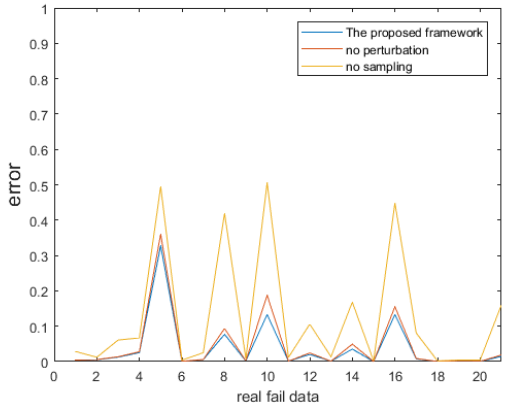


그림 5. 제안된 모델과의 분류 오류 비교

Fig. 5. Comparison of classification error

표 3. 각 분류모델의 혼동행렬

Table 3. Confusion Matrix of each classification model

(a) Proposed model

	Predicted Class		
Actual Class	Value	Class=1(fail)	Class=-1(pass)
	Class=1	80	4
	Class=-1	0	503

(b) k-NN

	Predicted Class		
Actual Class	Value	Class=1(fail)	Class=-1(pass)
	Class=1	58	1
	Class=-1	98	311

(c) Logistic Regression

	Predicted Class		
Actual Class	Value	Class=1(fail)	Class=-1(pass)
	Class=1	59	0
	Class=-1	137	272

(d) Naïve Bayes

	Predicted Class		
Actual Class	Value	Class=1(fail)	Class=-1(pass)
	Class=1	44	15
	Class=-1	144	265

(e) Decision Tree

	Predicted Class		
Actual Class	Value	Class=1(fail)	Class=-1(pass)
	Class=1	59	0
	Class=-1	66	343

표 4. 분류모델의 성능평가 비교

Table 4. Comparisons among Classification models

	k-NN	Logistic Regression	Naïve Bayes	Decision Tree	Proposed Model*
TPR	0.983	1.0	0.746	1.0	0.952
FPR	0.24	0.335	0.352	0.161	0.0
Precision	0.372	0.301	0.234	0.472	1.0
F-measure	0.54	0.463	0.356	0.641	0.976

5. 결론

반도체 공정에서 수율 향상을 위한 이상감지 및 분류 문제는 지속적으로 연구되어 오던 이슈이며, 다양한 기계학습방법들을 적용한 분류성능 향상이 시도되어 왔다. 대부분의 이상감지 및 분류 문제의 목표가 특징추출 부분에만 초점을 맞추어 온 반면, 본 연구는 누락된 값을 고려하고, 이를 통한 효과적인 소수 그룹의 데이터 생성에 초점을 맞추었다.

즉, GAN으로 소수그룹의 샘플 생성 과정에서 결측치 처리과정을 도입하여 실제 데이터와 유사한 가상데이터를 생성하고, 생성된 가상 불합격데이터를 분류모델의 학습과정에서 훈련데이터로 사용함으로써 생성된 데이터의 유효성을 검증하였다. 제시된 프레임웍이 기존의 프레임웍 보다 우수함을 증명하기 위해서 같은 데이터 셋을 사용한 기존 결과들과 혼동행렬을 통한 성능지표를 통해 평가를 수행하여 본 연구에서 제시한 방법론이 누락된 값을 고려한 오버샘플링 방법의 측면에서 효과적인 프레임웍을 제시하였음을 확인하였다.

본 연구는 반도체 공정뿐만 아니라 다른 여러 제조 공정에서 발생하는 불합격을 분류하기에 유효한 방법으로 활용될 수 있다. 향후 연구로는 다중 클래스에서의 불균형데이터 문제를 위한 프레임웍 등이 기대된다.

References

[1] J. Son, J. Ko and C. Kim, "Feature Based Decision Tree Model for Fault Detection and Classification of Semiconductor Process," *IE interfaces*, vol. 22, no. 2, pp. 126-134, 2009.

[2] S. Seo, Y. Jeon , J. Lee, H. Jung and J. Kim, "An Over-sampling Method based on Generative Adversarial Networks for Effective Classification of Imbalanced Big Data," in *Korean Institute of Information Scientists and Engineers*, 2017.

[3] K. Kim, B. Zhang and H. Jang, "Oversampling Based Ensemble Learning Methods for Imbalanced Data," *KIISE Transactions on Computing Practices*, vol. 20, no. 10, pp. 49-554, 2014.

[4] M. Randolph-Gips,member IEEE, "A New Neural Network to Process Missing Data without Imputation," *Seventh International Conference on Machine Learning and Application*, 2008.

[5] E.-L. Silva-Ramírez, R. Pino-Mejias, M. López-Coello and M.-D. Cubiles-de-la-Vega, "Missing value imputation on missing completely at random data using multilayer perceptrons," *Neural Networks*, vol. 24, no. 1, pp. 121-129, 2011.

[6] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, 2013.

[7] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of Translational Medicine*, vol. 4(1), no. 9, 2016.

[8] M. McCann and A. Johnston, "UCI Machine Learning Repository," [Online]. Available: http://archive.ics.uci.edu/ml/datasets/secom.

[9] B. E. Goodlin, D. S. Boning and H. H. Sawin, "Simultaneous Fault Detection and Classification For Semiconductor Manufacturing Tools," *Journal of the Electrochemical Society*, vol. 150, no. 12, pp. G778-G784, 20.

[10] K. Lakshminarayan, S. A. Harp and T. Samad, "Imputation of Missing Data in Industrial Databases," *Applied Intelligence*, vol. 11, pp. 259-275, 1999.

[11] J. Kaiser, "Dealing with Missing Values in Data," *Journal of Systems Integration*, vol. 5, no. 1, 2014.

[12] Y. C. Yuan, "Mutiple Imutation for Missing Data: Concepts and New Development," 2005.

[13] K. Kerdprasop and N. Kerdprasop, "A Data Mining Approach to Automate Fault Detection Model Development in the Semiconductor Manufacturing Process," *International Journal of Mechanics*, vol. 5, no. 4, 2011.

[14] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221-232, 2016.

[15] W. Nam and S. Kim, "A Prediction of Wafer Yield Using Product Fabrication Virtual Metrology Process Parameters in Semiconductor Manufacturing," *Journal of the Korean Institute of Industrial Engineers*, vol. 41, no. 6, pp. 572-578, 2015.

[16] Kittisak Kerdprasop and Nittaya Kerdprasop, Member, IAENG, "Feature Selection and Boosting Techniques to Improve Fault Detection Accuracy in the Semiconductor Manufacturing Process," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, 2011.

[17] D. Moldovan, T. Cioara, I. Anghel and I. Salomie, "Machine Learning for Sensor-Based Manufacturing Process," *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 147-154, 2017.

[18] S. Munirathinam and B. Ramadoss, "Predictive Models for Equipment Fault Detection in the Semiconductor Manufacturing Process," *IACSIT International Journal of Engineering and Technology*, vol. 8, no. 4, pp. 273-285, 2016.

[19] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with Applications*, vol. 91, pp. 464-471, 2018.

[20] 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, pp. 345-354, 2007.

[21] Y. Lee and J. Hwang, "Application of Deep Neural Network and Generative Adversarial Network to Industrial Maintenance: A Case Study of Induction Motor Fault Detection," in *IEEE International Conference on Big Data*, 2017.

[22] J. Kim, Y. Han and J. Lee, "Data Imbalance Problem solving for SMOTE Based Oversampling: Study on Fault Detection Prediction Model in Semiconductor Manufacturing Process," *Advanced Science and Technology Letters*, vol. 133, pp. 79-84, 2016.

[23] Y. LeCun, L. Bottou, B. B. Orr and K. -R. Müller, "Efficient Backprop," in *Neural Networks: Tricks of the Trade, Berlin, Heidelberg, Springer*, 1998, pp. 9-50.

[24] A. Iserles , Acta Numerica 2005: Volume 14, New York: the Press Syndicate of the University of Cambridge, 2005

저 자 소 개

김희수(Heesoo Kim)

2017년~현재 : 금오공과대학교 산업공학과 석사과정

2017년 : 금오공과대학교 산업공학부 공학사

관심분야 : Machine Learning, Deep Learning, Data Mining

Phone : +82-54-478-7681

E-mail : heees@kumoh.ac.kr

이현수(Hyunsoo Lee)

2011년~현재 : 금오공과대학교 산업공학부 부교수

2010년 : Texas A&M University 산업시스템공학과 박사

2002년 : POSTECH 산업공학과 공학 석사

관심분야 : Nonlinear Control and Optimization, Machine Learning, Virtual Manufacturing, Data Engineering

Phone : +82-54-478-7661

E-mail : hsl@kumoh.ac.kr