Journal of the Korea Academia-Industrial cooperation Society
Vol. 20, No. 9 pp. 551-557, 2019

https://doi.org/10.5762/KAIS.2019.20.9.551
ISSN 1975-4701 / eISSN 2288-4688

이상진단 성능에 미치는 변수선택과 추정방법의 영향

조현우
대구대학교 신소재에너지공학과

Effect of Different Variable Selection and Estimation Methods on Performance of Fault Diagnosis

Hyun-Woo Cho
Department of Materials-Energy Science and Engineering, Daegu University

요 약 생산 공정에서 발생하는 비정상적인 이상 (fault)의 진단 (diagnosis)은 고품질의 제품을 생산함에 있어 필수적이라 할 수 있다. 회분식 공정 (batch process)과 같이 부가가치가 큰 반도체나 의약품 등의 첨단 제품을 생산하는 공정에서는 더욱 실시간 진단의 역할이 커지고 있다. 본 연구에서는 회분식 공정으로부터 얻은 측정 데이터와 비선형 분류 (nonlinear classification)에 기초한 실시간 이상 진단 체계에 있어서 변수선택과 미래값 추정 기법이 진단 성능에 미치는 영향을 평가한다. 공정 변수 중 진단에 필수적이며 기여도가 높은 변수만을 선택하여 진단 모델 (diagnosis model)을 구성함으로써 진단 성능의 향상을 기대할 수 있다. 본 연구에서는 여러 변수선택 (variable selection) 기법들의 진단 성능을 비교 평가한다. 또한, 현재 진행 중인 회분식 조업 데이터는 종료되기 이전에는 진단에 필요한 전체 데이터를 얻을 수 없으므로 현재 시점에서 측정되지 못한 미래 측정값 (future observations)이 추정되어야 한다. 미래값 추정방법들의 선택이 변수선택과 분류기반 진단 관점에서 진단 성능에 어떻게 영향을 주는지 평가한다. 폴리염화비닐 회분식 공정에 대한 사례 연구를 수행하여 최적의 변수선택과 미래값 추정방법을 도출하였다. 변수선택 방법에 따라 최대 21.9%와 13.3%의 성능 향상을 보였으며 미래값 추정방법에 따라서는 최대 25.8%와 15.2% 향상됨을 알 수 있었다.

Abstract Diagnosis of abnormal faults is essential for producing high quality products. The role of real-time diagnosis is quite increasing in the batch processes of producing high value-added products such as semiconductors, pharmaceuticals, and so forth. In this study, we evaluate the effect of variable selection and future-value estimation techniques on the performance of the diagnosis system, which is based on nonlinear classification and measurement data. The diagnostic performance can be improved by selecting only the variables that are important and have high contribution for diagnosis. Thus, the diagnostic performance of several variable selection techniques is compared and evaluated. In addition, missing data of a new batch, called future observations, should be estimated because the full data of a new batch is not available before the end of the cycle. In this work the use of different estimation techniques is analyzed. A case study on the polyvinyl chloride batch process was carried out so that optimal variable selection and estimation methods were obtained: maximum 21.9% and 13.3% improvement by variable selection and maximum 25.8% and 15.2% improvement by estimation methods.

Keywords : Fault, Diagnosis, Batch Process, Classification, Variable Selection, Estimation

This research was supported by the Daegu University Research Grant, 2016.
*Corresponding Author : Hyun-Woo Cho(Daegu Univ.)
email: hwcho@daegu.ac.kr
Received April 29, 2019 Revised August 26, 2019
Accepted September 6, 2019 Published September 30, 2019

1. Introduction

생산 공정에서 발생하는 비정상적인 이상 (fault) 상황에 대한 신뢰성 있는 진단 (diagnosis)은 고품질 제품을 생산하기 위해서는 필수적이다. 특히 부가가치가 큰 반도체나 의약품 등의 첨단 제품을 생산하는 회분식 공정 (batch process)에서 이상 진단의 역할은 크다고 할 수 있다[1]. 일반적으로 회분식 공정의 조업은 원료 물질이 사전 정의된 순서와 양만큼 충전되어 시작되며 제어된 조건 하에서 이들이 처리되어 최종적으로 완성된 제품이 배출됨으로써 하나의 회분식 조업 사이클이 완성된다. 이러한 회분식 공정은 공정 변수의 궤적 (process variable's trajectory)들이 최소한의 편차로 유지되며 최종적으로 균일한 고품질의 제품을 얻었을 때 성공적인 조업으로 간주된다. 회분식 공정은 연속 공정 (continuous process)과는 달리 일정 시간 동안에만 조업되고 비선형성이 강한 공정 거동과 측정 데이터 (measurement data)등의 특성을 가진다. 이로 인하여 실시간으로 공정 이상을 진단하는 체계를 회분식 공정에 적용하는 것이 상대적으로 어려울 수 있다. 이러한 문제를 해결하기 위하여 multiway PCA (MPCA: multiway principal component analysis) 기법이 회분식 공정에 적용된 이후 support vector machines (SVM), 커널 PCA (KPCA), 커널 partial least squares (KPLS), 커널 Fisher discriminant analysis (KFDA)등의 다양한 비선형 다변량 통계적 기법들이 널리 활용되어 왔다[1-4].

본 연구에서는 비선형 분류 (nonlinear classification) 기법의 하나인 KFDA에 기반한 회분식 공정의 실시간 이상 진단 체계에 있어서 변수선택 (variable or feature selection)과 미래값 추정 (estimation of future observations) 기법이 진단 성능에 미치는 영향을 다룬다. 측정 데이터 중 전체 변수를 사용하는 대신 이상 진단에 필수적이며 기여도가 높은 변수만을 선별하여 진단 모델 (diagnosis model)을 구성함으로써 진단 성능의 향상을 기대할 수 있다. 본 연구에서는 PCA, PLS, SVM에 기반한 3가지 변수선택 기법들과 변수선택을 하지 않는 경우 이렇게 총 4가지 경우의 진단 성능을 비교 평가한다. 한편, 회분식 공정의 이상 진단을 실시간으로 수행하는 경우, 현재 진행 중인 조업이 종료되기 이전에는 진단에 필요한 전체 데이터를 얻을 수 없기 때문에 현재 시점에서 측정되지 못한 측정 데이터인 미래 측정값이 추정되어야만 한다. 본 연구에서는 다른 2가지 미래값 추정 방법들의 선택이 변수선택과 분류기반 KFDA 진단 관점에서 진단 성능에 어떻게 영향을 주는지 평가하고자 한다. 이를 위하여 우선 회분식 공정 데이터의 특성과 앞에서 언급된 방법론들을 살펴보고 폴리염화비닐 (PVC: polyvinyl chloride) 회분식 공정의 데이터를 활용한 사례의 진단 결과를 제시하며 변수선택과 미래값 추정방법론에 따른 영향을 알아본다.

2. 방법론

회분식 공정의 측정 데이터는 Fig. 1에 나타낸 것과 같이 연속 공정의 2차원 데이터와는 달리 three-way array 형태를 갖고 있다. 하나의 회분식 조업 (batch run)은 조업 기간 중 K 샘플링 타임마다 J개 공정 변수 (j=1, 2, … , J)가 동일하게 측정되게 된다. 이와 동일한 형태의 회분식 측정 데이터가 총 I개 회분식 조업 (i=1, 2, … , I)별로 존재하기 때문에 Fig. 1에서 보이는 three-way array 의 형태를 갖게 된다. 이러한 형태의 회분식 측정 데이터를 분석하기 위해서 unfolding이라는 데이터 처리를 활용하게 된다. Fig. 1에서 보는 바와 같이 구체적으로 unfolding이란 의 (I×J) 부분을 시간 순서대로 옆으로 위치시키는 데이터 처리를 의미한다. 이렇게 unfolding 함으로써 데이터 분석 과정에서 대신 펼쳐진 (unfolded) 이차원 매트릭스 (I×JK)를 사용할 수 있게 되어 기존 다변량 통계적 기법들을 제한 없이 적용할 수 있게 된다[3].

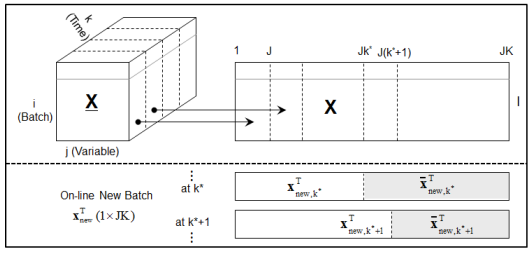

Fig. 1. Characteristics of batch data

Unfolding을 통해 과거에 조업이 종료된 데이터를 분석하는 것은 어려움이 없으나 현재 조업 중인 신규 조업 (new batch) 데이터를 실시간으로 분석할 때에는 미래값 추정이 필요하다. Fig. 1의 아래에서 보이는 것처럼, 특정 샘플링 타임 k^*에서 펼쳐진 신규 조업 데이터 x_{new}^{T} (1×JK)는 이미 측정된 데이터 부분 x_{new,k^*}^{T}과 미측정된

데이터 부분 $\bar{x}_{new,k^*}^T$ 즉 미래값 (future observations)으로 구성되는데 이 미래값이 추정되어야 하는 것이다. 이와 동일하게 다음 샘플링 타임 k^*+1에서는 이미 측정된 x_{new,k^*+1}^T와는 다르게 미측정된 미래값 $\bar{x}_{new,k^*+1}^T$은 추정값이 필요하다. 자주 사용되는 current deviation 미래값 추정은 현재 시점에서 평균 궤적 (mean trajectory)과의 현재 편차 (current deviation)를 고려하는 반면에 fault library 미래값 추정은 현재 시점에서 과거 관측된 조업 중 가장 유사한 궤적을 선택하여 미래값을 추정하게 된다[5].

SVM은 주어진 데이터로부터 분류 및 회귀 규칙을 학습하는 기법이다. 데이터는 우선 high-dimensional feature space로 매핑되는데 여기에서 최대 마진 (maximum margin)을 갖도록 최적의 결정 함수 (optimal decision function)가 주어진다. 이 결정 함수는 다음의 부등식을 만족하여야 한다[6].

$$y_i(\mathbf{w}\Phi(\mathbf{x}_i)+b)-1\geq 0\ \forall_i \tag{1}$$

이러한 결정 함수를 통해 아래의 dual problem으로 변환되는데 SVM을 데이터로 학습한다는 것은 곧 주어진 커널 함수 조건에서 αi, b, and support vectors를 찾는 것이다.

$$L_d=\sum\alpha_i-1/2\sum\alpha_i\alpha_j y_i y_j\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j) \tag{2}$$

SVM을 통한 변수선택 방법 중 반복적 변수 제거 (recursive feature elimination) 절차와 결합한 SVM-RFE는 주어진 비용 함수 (cost function)에 대한 민감도 분석 (sensitivity analysis)을 실행한다. 데이터의 클래스 레이블 $y=[y_1,...,y_l]^T$을 가진 $X_{all}=[x_1,...,x_l]$ 학습 샘플에 대하여 다음과 같이 비용 함수를 정의하는데[6]

$$J=(1/2)\alpha^T H\alpha-\alpha^T e \tag{3}$$

여기서 e는 1로 구성된 벡터, α는 Lagrange multiplier, $H_{hk}=y_h y_k K(x_h,x_k)$이다. 제거되는 i번째 변수로 인해 생긴 J의 변화량을 얻기 위해 α은 일정하다고 가정하면

$$H(-i)_{hk}=y_h y_k K(x_h(-i),x_k(-i)) \tag{4}$$

여기서 $(-i)$는 i번째 변수가 제거되었음을 의미한다. 그 결과 민감도 함수는 다음으로 주어진다.

$$\begin{aligned}DJ(i)&=J-J(-i)\\&=(1/2)\alpha^T H\alpha-(1/2)\alpha^T H(-i)\alpha.\end{aligned} \tag{5}$$

SVM-RFE 알고리즘은 s가 empty array가 될 때까지 아래의 절차를 반복하여 진행된다[6].

(1) 신규 샘플을 구성한다 $X=X_{all}(:,s)$

(2) SVM을 학습하여 α를 구한다

(3) 아래와 같이 ranking criterion을 구한다
$DJ(i)=(1/2)\alpha^T H\alpha-(1/2)\alpha^T H(-i)\alpha$

(4) $f=\arg\min_i DJ(i)$을 만족하는 변수 f를 찾는다

(5) r을 업데이트하고 s에서 변수 f를 제거한다
$r=[s(f),r], s=s-s(f).$

Linear discriminant analysis의 비선형 기법인 KFDA는 비선형 특성 공간 (nonlinear feature spaces)에서 선형 discriminant analysis를 실행하게 된다. 여기서 비선형 discriminant vectors는 다음을 최대화하여 얻을 수 있는데[1]

$$J^{\Phi}(\boldsymbol{\psi})=\frac{\boldsymbol{\psi}^{\mathrm{T}}\mathbf{S}_b^{\Phi}\boldsymbol{\psi}}{\boldsymbol{\psi}^{\mathrm{T}}\mathbf{S}_t^{\Phi}\boldsymbol{\psi}} \tag{6}$$

여기서 $\mathbf{S}_b^{\Phi}$과 $\mathbf{S}_t^{\Phi}$은 클래스 간 공분산 (between class covariance)과 전체 공분산 (total covariance)을 각각 의미한다. 최적인 discriminant vectors는 아래 식에서 구할 수 있다.

$$\mathbf{S}_b^{\Phi}\boldsymbol{\psi}=\lambda\mathbf{S}_t^{\Phi}\boldsymbol{\psi} \tag{7}$$

또한 아래 식을 만족하는 계수 b_k가 존재하는데

$$\boldsymbol{\psi}=\sum_{k=1}^{M}b_k\Phi(\mathbf{x}_k)=\mathbf{H}\boldsymbol{\alpha}, \tag{8}$$

여기서 $\mathbf{H}=[\Phi(\mathbf{x}_1),\ldots,\Phi(\mathbf{x}_M)]$, $\boldsymbol{\alpha}=(b_1,\ldots,b_M)^{\mathrm{T}}$. KFDA에서는 비선형 매핑 $\Phi(x)$을 통해 입력 데이터를 새로운 공간으로 투영한 후 선형 discriminant analysis를 적용한다.

3. 사례 결과

본 연구에서 제안된 진단 체계의 성능을 평가하기 위

하여 폴리염화비닐 회분식 공정에 대한 사례 연구를 진행하였다. 검증 대상 공정은 중합 반응기 (polymerization reactor), 환류 응축기 (reflux condenser), 교반기 (agitator) 및 냉각 재킷 (cooling jacket)으로 구성되어 있다. 총 11개의 공정 변수가 총 241 샘플링 타임에 온라인으로 자동 측정된다[5]. 여기서는 온도, 압력, 유속 (flow rate), 교반기 속도 (agitator speed) 등과 같은 조업 상태를 측정하는 공정 변수들을 의미한다. 특히 반응기 온도는 미리 정해진 궤적을 따르도록 조심스럽게 제어되어야 한다.

이러한 대상 공정으로부터 다섯 가지 공정 이상 클래스 (fault classes)별로 각각 일곱 개의 이상 조업 데이터 (fault batches)를 준비하여 이 데이터를 KFDA 분류 기반 진단 모델의 학습 데이터 (training data)로 사용한다. PVC 회분식 공정에서 241개 샘플링 타임에 온라인으로 측정하는 11개 공정 변수가 있기 때문에 결과적으로 unfolded training data Z(35 x 2651)를 얻게 된다. 35개 training data와는 별도로 진단 성능을 독립적으로 평가하기 위하여 이상 클래스별로 두 개씩 총 10개의 회분식 조업 데이터를 테스트 데이터로 사용한다.

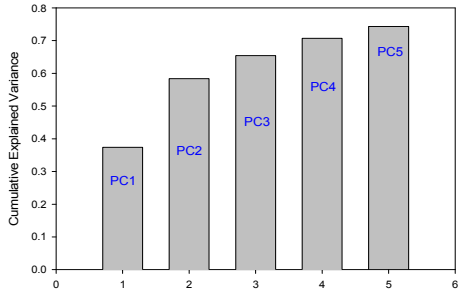

Fig. 2. PCA model for monitoring data

이상 진단의 전 단계로서 공정 이상의 발생을 감지하여 이후 진단 기능을 호출하는 역할을 가진 모니터링 모델은 반드시 필요하다. 본 연구에서는 총 44개의 이상 발생이 없는 PVC 정상 조업 데이터에 대해 multiway PCA 모델을 구성하였다. Fig. 2는 이렇게 구성한 multiway PCA 모니터링 모델에서 사용된 5개의 주성분 (principal component, PC)들이 전체 데이터 변화량을 얼마나 설명하고 있는지를 누적 %로 (cumulative explained variance) 제시하고 있다. 정의상 가장 많은 데이터 변화량을 설명하는 첫 번째 주성분 (PC1)이 37%를, 그리고 전체적으로 총변화량의 74.4%를 설명하고 있음을 알 수 있다.

이상 진단 모델 구성에 앞서 데이터의 전처리 (preprocessing)로서 orthogonal signal correction (OSC)을 실행하였다. OSC는 일반적으로 응답변수 (response variable) Y와 상관성이 없는 독립변수 (independent variable) X의 변화량을 없애주는 노이즈 필터링 (noise filtering) 효과를 주는 장점이 있다[7]. 본 연구의 OSC 필터링은 공정 이상 클래스를 분류하는 진단 체계이므로 여기에 맞게 OSC를 PLS-discriminant analysis (PLS-DA) 방식으로 적용하였다. 분류 문제에 적합한 PLS-DA 방식에서는 응답변수 Y에 공정 이상 클래스에 대한 멤버십 정보를 포함한 후 진행한다. 예를 들어 PLS-DA의 멤버십 행렬 Y의 첫 번째 행은 [1 0 0 0 0]이며, 이는 첫 번째 공정 이상 클래스에 속함을 의미합니다. 앞에서 언급한 PVC 회분식 공정의 unfolded training data에 OSC를 적용하여 (35x2,651) 크기의 전처리된 데이터를 얻었다.

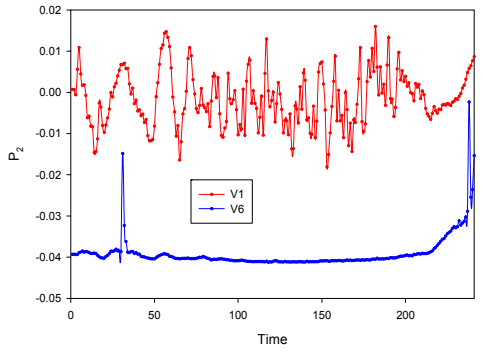

Fig. 3. Loading plots for two variables

다음 단계로 OSC 필터링된 데이터에 대하여 SVM-RFE 변수선택 알고리즘을 적용하여 이상 클래스 분류에 기여도가 높은 변수들을 선택하였다. 즉, 총 2,651개의 펼쳐진 변수들 중에서 분류 진단에 도움을 주는 변수를 선별하여 이들 변수들만으로 KFDA 진단 모델을 구성하게 된다. 또한 변수선택 기법의 비교를 목적으로 PCA와 variable importance (VIP)에 기반한 2가지 변수선택 기법을 테스트하였다. 첫 번째 기법은 PCA의 로딩 벡터 (loading vector) 정보를 변수선택에 활용한다. 즉, 개별 변수는 PCA의 축소된 공간 (reduced spaces)을 구성할 때 동일한 중요도를 갖지 않으며 따라서 PCA의 로딩 계수 (loading coefficients)는 축소된 공간에서의 변수 중요도를 의미하게 된다.

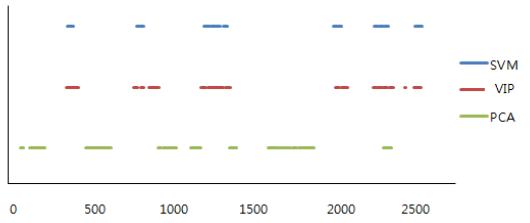

Fig. 4. Regions of variables selected by three methods

예를 들면, Fig. 3에서 보이듯이 아래 부분의 여섯 번째 변수 v6에 해당하는 로딩 계수들은 v1의 로딩 계수보다 절대값이 크므로 상대적으로 중요하다고 할 수 있다. 이를 수식화하면, j 번째 변수에 대한 PCA 변수선택 지수는 아래와 같이 주어지는데[2]

$$PCA_j = \sum_{i=1}^{A} |p_{ij}| EV_i \tag{9}$$

여기서 A는 retained PCs의 수 그리고 EV_i는 전체 변화량중에서 I번째 PC에 의해 설명되는 비율을 의미한다.

비교 대상인 다른 변수선택 기법은 PLS-DA 분석의 변수 중요도 VIP 정보를 활용하는 것이다. 앞의 OSC에서 언급했듯이 PLS-DA는 서로 다른 클래스의 샘플들을 최대로 분리시키는데 이를 위해 각 샘플에는 특정 클래스에 속하는지 여부에 따라 1 또는 0 값이 지정된 후 PLS를 적용하게 된다. 이 과정에서 PLS-DA를 적용하여 Y의 클래스 멤버십을 결정할 때 X의 어떤 변수들이 중요한 역할을 갖는지 결정할 수 있다. 이를 위하여 j 번째 변수에 대한 VIP 변수선택 지수 VIP_j는 아래와 같다[2].

$$VIP_j = \frac{N\sum_{i=1}^{A} w_{ij}^2 RSS_i}{RSS_T} \tag{10}$$

여기서 w_{ij}는 PLS-DA 계수 그리고 RSS는 explained residual sum of squares (%)를 의미한다. Fig. 4는 SVM, VIP, PCA에 기반한 세 가지 변수선택의 결과를 보여주는데 이 그림에서 선들은 데이터의 변수 중에서 선택된 변수들의 영역을 나타낸다. 전체 2,651 변수 중 SVM, VIP, PCA에 의해 각각 368, 571, 760개의 변수들이 선택되었는데, PCA가 가장 많은 변수를 선택한 반면 SVM은 가장 적은 변수 (전체 변수의 13.9%)를 선택하였다.

Table 1. Diagnosis results

	DSR_{20}/DSR_{end}(%)			
	Without	PCA	VIP	SVM
1A	75/85	70/81	80/89	85/93
1B	75/83	65/77	75/86	80/91
2A	85/91	80/88	90/94	90/94
2B	90/94	85/92	90/94	95/97
3A	80/88	70/82	85/90	85/91
3B	80/85	75/82	80/88	90/94
4A	85/92	75/81	90/94	95/97
4B	80/89	70/81	85/91	95/97
5A	80/89	75/85	85/92	90/94
5B	75/87	65/82	80/90	80/92
Avg	81/88	73/83	84/91	89/94

이러한 서로 다른 변수선택의 결과를 바탕으로 KFDA 이상 진단 모델을 각각 구성하고 10개의 테스트 데이터에 적용하여 진단 결과를 얻을 수 있었다. 본 연구에서는 PVC 회분식 공정에 대한 진단 성능 비교를 위해 위에서 언급한 3가지 변수선택 기법들 이외에 변수선택이 없이 전체 변수 모두를 진단 모델에 사용한 경우 ('without'로 표시)를 고려하여 총 4가지 경우의 진단 결과에 대해 비교하였다.

Table 1에서는 대상 공정인 폴리염화비닐 회분식 공정의 테스트 데이터인 10개 회분식 조업에 대한 진단 결과를 나타내었다. 공정 이상 클래스별 2개의 서로 다른 테스트 데이터를 구별하기 위해 숫자 뒤에 A 또는 B를 추가하였다. 예를 들어 첫 번째 이상 클래스의 경우 1A와 1B로 구분된다. 위의 표에서 진단 결과는 2가지 diagnosis success rate (DSR)의 값, 즉 DSR_{20}과 DSR_{end}로 주어진다. DSR은 정확하게 진단된 관측치의 비율로서 (%), DSR_{20}은 이상 감지 시점인 k*부터 20번째 샘플링 타임까지의 DSR을 나타내고 DSR_{end}는 k*부터 조업 종료까지의 DSR값으로 정의된다. 테스트 데이터

Table 2. Diagnosis results using current deviation

	DSR_{20}/DSR_{end}(%)			
	Without	PCA	VIP	SVM
1A	70/82	65/79	70/84	75/90
1B	65/82	60/77	75/87	75/90
2A	75/86	70/81	80/91	80/90
2B	80/89	75/83	80/91	85/92
3A	65/81	60/78	70/87	75/91
3B	65/81	50/75	65/85	70/89
4A	70/88	65/82	75/89	80/93
4B	70/86	60/80	75/87	85/95
5A	75/84	60/78	70/85	80/90
5B	65/82	55/76	80/88	75/88
Avg	70/84	62/79	74/87	78/91

1A의 예를 보면, 변수선택 없이 진단한 결과는 'without'에 제시되어 있는데 75%의 DSR_{20}과 85%의 DSR_{end} 값을 가지고 있다.

전반적인 진단 성능을 비교해보면 SVM의 변수선택 시 가장 좋은 진단 성능을 얻었으며, 표 하단에 표시된 평균 관점에서도 89%의 DSR_{20}과 94%의 DSR_{end} 가장 높은 평균값을 보여주고 있다. 반면에 73%의 DSR_{20}과 83%의 DSR_{end} 평균값을 가진 PCA 변수선택이 가장 좋지 않은 결과를 나타냈다. 특히 PCA 변수선택의 진단 결과가 변수선택이 없었던 'without'의 결과보다 더 저하되었다. 이것은 PCA분석이 데이터의 클래스 간 구분을 최대화시키기 보다는 단순히 데이터의 변화량을 설명하도록 축소된 공간 (reduced spaces)을 구성하는 특성에 기인한 것으로 판단된다. 앞에서 제시된 Fig. 4의 선택된 변수 영역을 다시 살펴보면, PCA의 결과가 나머지 SVM과 VIP의 결과와 중첩된 부분이 거의 없는 것을 알 수 있다. 이렇게 PCA 특성에 맞게 선택된 변수들로 진단 모델을 구성하였기 때문에 SVM과 VIP는 물론 변수선택이 없었던 경우보다 더 진단 성능이 저하된 것으로 보인다.

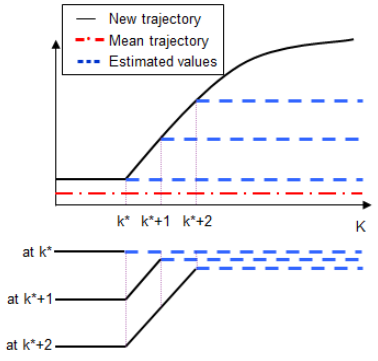

Fig. 5. Future estimation by current deviation

다음으로는 미래값 추정방법에 따른 진단 성능을 평가하기 위하여 Table 1에서 사용된 fault library 미래값 추정방법 대신 current deviation을 사용한 진단 결과를 Table 2와 같이 도출하였다. 우선 Table 1과 비교할 때 변수선택의 영향 관점에서 유사한 결과가 관측된다. 즉, SVM 변수선택 시 가장 좋은 진단 성능을 보였으며 (평균 89% DSR_{20}, 평균 94 DSR_{end}), 다음으로 VIP, without, PCA 순서를 보여준다. 한편, Table 2의 current deviation 미래값 추정방법을 사용한 진단 성능은 Table 1과 비교했을 때 변수선택에 상관없이 모든 테스트 데이터에서 fault library보다 감소한 것을 알 수 있다. 이는 평균 DSR_{20} 및 평균 DSR_{end} 측면에서도 그러한데, 여기서 주목할 점은 DSR_{end}에서의 성능 저하보다는 DSR_{20}의 그것이 상대적으로 더욱 크다는 것이다. 하나의 예로서 SVM의 경우를 살펴보면, 평균 DSR_{end}은 94%에서 91%로 감소한 반면 평균 DSR_{20}은 89%에서 78%로 대폭 감소한 것을 알 수 있다. 이와 같은 패턴은 다른 변수선택 방법들을 사용한 진단 결과에서도 볼 수 있다.

Current deviation에 의해 미래값이 추정될 때 이상 발생 초기의 DSR_{20}값이 저하되는 이유는 Fig. 5의 current deviation 특성으로 설명이 가능하다. 일반적으로 회분식 공정의 신규 조업 데이터가 평균 궤적 (mean trajectory)을 따를 경우 current deviation은 신뢰성 있는 미래 추정값을 준다. 그러나 그림에서 보이듯이 k* 이후 공정 이상이 발생하여 평균 궤적을 이탈하게 되면 current deviation에 의한 미래 추정값들은 실제 신규 데이터의 궤적을 빠르게 따라가지 못하게 되어 민감도가 떨어지게 된다. 이러한 특성으로 이상 발생 초기의 진단 결과치인 DSR_{20}값이 저하된다는 것은 이상 발생 직후 잘못 진단된 원인파악을 의미하며 이는 곧 회분식 공정의 정상화 지연으로 이어질 수 있어 주의가 필요하다.

4. 결론

본 연구에서는 비선형 분류 KFDA에 기반한 회분식 공정의 실시간 이상 진단 방법론을 폴리염화비닐 공정에 적용하였다. 그 과정에서 데이터의 크기를 줄이며 분류에 필요한 정보만을 선별하여 진단 성능을 향상시킬 수 있는 변수선택 방법들을 진단 성능 측면에서 평가하였다. SVM과 VIP 변수선택에 의한 진단 결과는 향상된 반면에 PCA 변수선택은 기법의 특성으로 인하여 오히려 변수선택이 없는 경우보다 성능이 저하되었다. 추가적으로 어떤 미래값 추정방법을 사용하는지에 따라 진단 성능이 어떻게 영향을 받는지 살펴보았다. 이상 발생 초기에 정확한 진단 결과를 얻어 빠르게 이상의 원인을 제거하기 위해서는 대상 공정에 적합한 미래값 추정방법을 선정해야 한다.

References

[1] Z. Ge, "Process data analytics via probabilistic latent

variable models: a tutorial review", *Industrial and Engineering Chemistry Research*, Vol.57, No.38, pp.12646-12661, 2018.
DOI: https://doi.org/10.1021/acs.iecr.8b02913

[2] K. Tidriri, N. Chatti, S. Verron, T. Tiplica, "Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges", *Annual Reviews in Control*, Vol.42, pp.63-81, 2016.
DOI: https://doi.org/10.1016/j.arcontrol.2016.09.008

[3] H. Rostami, J. Blue, C. Yugma, "Automatic equipment fault fingerprint extraction for the fault diagnostic on the batch process data", *Applied Soft Computing*, Vol.68, pp.972-989, 2018.
DOI: https://doi.org/10.1016/j.asoc.2017.10.029

[4] S. Bersimis, S. Psarakis, J. Panaretos, "Multivariate Statistical Process Control Charts: An Overview", *Quality and Reliability Engineering International*, Vol.23, pp.517-543, 2007.
DOI: https://doi.org/10.1002/qre.829

[5] H. Cho, K. Kim, "A method for predicting future observations in the monitoring of a batch process", *Journal of Quality Technology*, Vol.35, pp.59-69, 2003.
DOI: https://doi.org/10.1080/00224065.2003.11980191

[6] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines", *Machine Learning*, Vol.46, pp.389-422, 2002.
DOI: https://doi.org/10.1023/A:1012487302797

[7] S. Wold, H. Antti, F. Lindgren, J. Ohman, "Orthogonal signal correction of near-infrared spectra", Chemometrics and Intelligent Laboratory Systems, Vo.44, pp.175-185, 1998.
DOI: https://doi.org/10.1016/S0169-7439(98)00109-9

조 현 우(Hyun-Woo Cho) [정회원]

- 2003년 8월 : 포항공과대학교 기계산업공학부 (공학박사)
- 2003년 8월 ~ 2007년 8월 : 포항공과대학교, 조지아텍, 테네시주립대 연구원
- 2007년 9월 ~ 2011년 2월 : 삼성전자, 삼성디스플레이 책임연구원
- 2011년 3월 ~ 현재 : 대구대학교 신소재에너지공학과 교수

〈관심분야〉
공정모니터링, 데이터 마이닝, 신재생에너지