다변량 시계열데이터 집단 상호간의 주성분 정보 비교를 통한 이상 예측

Fault Detection of Multivariate Time Series Data by Between - Group Comparison of Principle Components Information

저자
(Authors)

박찬영, 김창욱

출처
(Source)

한국경영과학회 학술대회논문집 , 2013.5, 1177-1184(8 pages)

발행처
(Publisher)

한국경영과학회
The Korean Operations Research and Management Science Society

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07171161

APA Style

박찬영, 김창욱 (2013). 다변량 시계열데이터 집단 상호간의 주성분 정보 비교를 통한 이상 예측. 한국경영과학회 학술대회논문집, 1177-1184

다변량 시계열데이터 집단 상호간의 주성분 정보 비교를 통한 이상 예측

Fault Detection of Multivariate Time Series Data by Between-Group Comparison of Principle Components Information

박찬영, 김창욱
연세대학교 정보산업공학과
{vici2014, kimco}@yonsei.ac.kr

Abstract

모니터링 시스템에서 계측되는 다변량 시계열 데이터를 분석하여 시스템의 상태가 변화했는지를 탐지하는 현재까지의 방법은 각 변수별로 시계열 데이터의 특징을 단일 값으로 요약하고 요약된 변수 값을 분류모형에 입력하여 이상 유무를 추정하는 방식이었다. 그러나 이 방법은 시계열 데이터를 요약하는 과정에서 데이터의 구조적 특징이 손실되는 단점이 있다. 본 논문은 시계열 데이터를 모두 사용하는 이상탐지 방법을 제안한다. 템플릿 시계열 데이터 집단과 테스트 시계열 데이터 집단이 주어졌을 때 각 데이터 군을 주성분 분석을 통해서 주성분 공간내의 score 분포로 표현하고 두 분포의 특징(주성분간의 사이각, 고유값의 비율, score 평균값의 비율)을 비교해서 거리 척도로 표현하고 Hampel 값을 이용하여 이상 여부를 탐지한다. 세 종류의 다변량 시계열 데이터를 대상으로 실험을 한 결과 본 연구에서 제안한 방법은 우수한 이상탐지 성능을 보였다.

Keywords: 다변량 시계열 데이터; 이상 예측; 주성분 분석; 거리척도; 반도체 공정.

1. 서론

최근 센서 및 정보 기술의 발달로 반도체 생산 설비, 주식 시장, 지구 대기 등, 관리의 대상이 되는 시스템의 실시간 모니터링이 가능해 졌다. 이런 모니터링 시스템의 특징으로는 시스템의 상태를 대변하는 각종 변수가 시계열 형태의 데이터로 수집되며 데이터 간에는 서로 상관관계가 있다는 점이다. 본 연구의 목적은 다변량 시계열 데이터로부터 시스템의 상태가 정상인지 아니면 이상인지를 판별하는 통계적 분류(Statistical classification) 기법을 제안하는 것이다. 현재 다변량 시계열 데이터의 이상 탐지는 시스템의 관리 차원에서 매우 중요한 의제로 대두되고 있다. 예를 들어 반도체 생상 공정에서는 센서로부터 수집되는 설비변수의 시계열 데이터를 분석하여 제품의 이상 여부를 미리 예측할 수 있어 수율 향상에 도움을 준다[4]. 또한 주식시장에서는 일정기간동안의 각종 경제 지표의 시계열 데이터를 분석하여 향후 종합주가지수의 상승여부를 예측해 볼 수 있으며[15], 지구 대기시스템에서는 각종 기후에 영향을 주는 인자의 실시간 모니터링을 통해서 지구 온난화의 지속여부를 예측할 수 있다[12].

현재까지의 다변량 시계열 데이터의 이상 예측을 위한 방법은 템플릿(Template) 기반 접근법을 이용한다. 이 방법은 예측하고자 하는 테스트 데이터의 시계열 패턴과 기준이 되는 템플릿(레퍼런스라고도 함)의 패턴을 비교하여 테스트 데이터의 이상 여부를 판단한다. 예를 들어 반도체 설비에서 한 개의 웨이퍼를 가공하는 시간 t동안 n 개의 변수가 모니터링 된다면 이 시스템은 매번 웨이퍼 가공이 끝난 후 길이 t를 갖는 시계열 데이터가 n 개 존재하는 시계열 데이터 집단(Group)을 생성한다. 가장 이상적인 가공결과를 보인 데이터 집단을 템플릿으로 설정한다. 그리고 새로 가공이 끝난 데이터 집단을 테스트 집단이라 한다. 템플릿 기반 접근법은 각 변수별 테스트 시계열 데이터마다 대응되는 템플릿 데이터와 패턴을 비교하여 단일 값으로 비교치를 표현한다. 대표적인 방법으로는 시계열 데이터의 평균을 비교하는 방법[14], 구조적 특징(평균,

분산, 최대/최소, 이상치 개수 등)을 비교하는 방법[2][9], 그리고 시계열 패턴을 비교하는 방법[8][10]이 존재한다. 각 변수별로 템플릿 데이터와 비교하여 단일 값으로 요약되면 n개의 요약 값을 Hotelling's T^2[5][8], 인공신경망[7], 의사결정나무[1], Support vector machine[14], k-nearest neighbor[6] 등의 다변량 분류 모형에 입력하여 이상여부를 판별한다. 그러나 이와 같은 절차는 첫째, 분류 모형은 시계열 데이터를 수용할 수 없으므로 전체 시계열 데이터를 사용하지 않고 n개의 요약된 변수 값으로 만드는 과정에서 데이터의 구조적 특징이 손실되는 단점이 있다. 둘째, 분류 모형을 학습시키는 과정에서 데이터의 제약성(일반적으로 반도체 공정에서는 정상 데이터 대비 이상 데이터의 비율이 매우 낮다) 때문에 수립된 분류 모형이 과잉맞춤(Over-fitting)될 가능성이 있다.

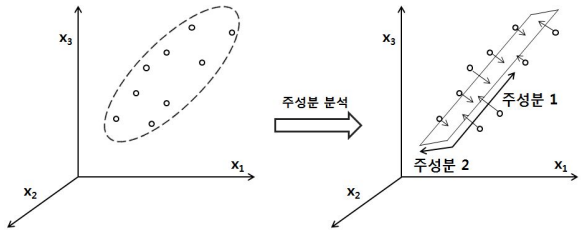
본 논문은 시계열 데이터를 모두 사용하는 이상탐지 방법을 제안한다. 템플릿 시계열 데이터 집단과 테스트 시계열 데이터 집단이 주어졌을 때 각 데이터 군을 주성분 분석을 통해서 주성분 공간내의 score 분포로 표현하고 두 분포의 특징(주성분간의 사이각, 고유값의 비율, score 평균값의 비율)을 비교해서 이상 여부를 탐지한다.

논문의 구성은 다음과 같다. 2장은 주성분간의 사이각, 고유값의 비율, score 평균값의 비율을 고려한 세 가지 거리 척도를 제안한다. 3장은 거리척도로부터 얻어지는 값을 이용하여 이상 여부를 결정하는 모델을 소개한다. 4장은 실험 결과를 소개하며, 마지막으로 5장은 결론 및 추후 연구방향에 대해 논한다.

2. 주성분 분석과 거리척도

2.1 주성분 분석

주성분 분석(Principal components analysis)이란 다변량 변수 자료를 분석하는 기법 중 하나로 원 변수 축 상에 분포된 자료를 선형 변환하여 주성분(Principal component)이라 불리는 잠재변수 축 상에 표현하는 기법이다. 주성분은 원 변수들의 선형결합으로 만들어지는데 자료의 분산이 가장 큰 방향이 제 1 주성분이 되고 그 다음 분산이 큰 방향이 제 2 주성분이 되며, 이런 과정을 반복하면 원 변수의 수만큼 주성분이 만들어진다. 그러나 일반적으로 분산이 작은 방향은 자료를 설명하는데 큰 도움이 안 되므로 주성분으로 간주하지 않는다. 주성분은 좌표축이기 때문에 서로 독립이어야 하며 각 주성분이 설명하는 자료의 분산을 고유값(Eigenvalue)라고 한다.



[그림 1] 주성분 분석에 의한 3차원 데이터의 2차원 데이터로의 변환

[그림 1]에서 보듯이 기존의 3차원의 데이터가 세 개의 원 변수 x_1, x_2, x_3 축으로 표현되었는데 주성분 분석을 통하면 두 개의 주성분 축(주성분 1, 주성분 2)으로 2차원의 데이터로 표현이 가능하다. 이 경우 데이터를 표현하는 변수의 수가 축소되었다. [그림 1]의 오른쪽 그림은 원 데이터 값을 두 개의 축 '주성분 1', '주성분 2' 의 좌표위에 표현한 그림이며 이때 데이터의 좌표값을 score 라고 하고, score 값은 왼쪽 그림의 원 데이터를 선형변환하여 얻은 2차원의 데이터가 된다. 주성분은 원 변수들의 선형결합으로 만들어지는데 이 때 원 변수가 주성분에 기여하는 계수를 loading 이라고 한다. 원 변수의 loading 값이 크다는 의미는 그 변수가 해당 주성분에 영향을 많이 끼친다는 의미이다.

본 연구에서는 템플릿 시계열 데이터 집단과 테스트 시계열 데이터 집단이 주어졌을 때 각 데이터 군을 주성분 분석을 통해서 주성분 공간내의 score 분포로 표현한다. 그리고 다음과 같은 세 가지 거리 척도를 이용해서 두 집단간의 유사성을 측정하고 테스트 시계열 데이터 집단의 이상 여부를 판별한다.

(1) 거리척도 1: 두 데이터 집단이 만들어낸 주성분들의 사이각을 쌍대 비교(Pairwise comparison)하며 얼마나 데이터 공간상에서 두 집단의 주성분 축이 유사한지를 측정한다.

(2) 거리척도 2: 두 데이터 집단의 고유값의 비율을 측정하여 얼마나 두 집단의 모양이 유사한지를 비교한다.

(3) 거리척도 3: 주성분 축상에 분포한 각 데이터 집단의 score 값의 평균을 비교하여 두 분포의 중심 위치가 얼마나 유사한지를 측정한다.

다음 절부터는 각각의 거리척도의 정의와 왜 이 척도가 필요한지를 설명한다.

2.2 거리척도 1

첫 번째 거리 척도는 템플릿 데이터 집단과 테스트 데이터 집단을 각각 주성분 분석하여 주성분 끼리의 사이각을 비교하여 계산

한다. Krzanowski[11] 는 이와 유사한 방법을 제안한 바 있다. 그는 주성분 분석을 이용한 두 그룹간의 차이를 측정하는 척도로 주성분들이 생성하는 초평면의 사이각으로 계산하는 식을 제안하였다. 그는 비교하고자 하는 두 그룹의 데이터를 각각 주성분 분석하여 각각의 주성분을 행렬로 표현하였다. 그리고 그 행렬들의 곱으로 얻어지는 새로운 행렬의 고유값을 크기순으로 $\lambda_1, \lambda_2, \ldots, \lambda_k$ 로 놓고 다음과 같은 두 그룹의 차이를 측정하는 유사도를 제안하였다.

$$\text{유사도} = \sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} \sum_{j=1}^{k} \cos^2 \theta_{ij} \qquad (1)$$

여기서 k는 비교하고자 하는 주성분의 수이고 θ_{ij} 는 첫 번째 그룹의 i 번째 주성분과 두 번째 그룹의 j 번째 주성분간의 사이각이다.

Krzanowski 의 방법은 결국 두 그룹의 주성분으로 생성되는 초평면들간의 사이각을 이용하는 방법인데 실제 적용시에 다음과 같은 문제점이 있다. [표 1]에서 보듯이 k=2이고 비교하고자하는 두 그룹의 데이터의 주성분이 일치하는 경우와 일치하지 않는 경우에 대해 유사도값이 같게 나오는 모순이 있다.

[표 1] Krzanowski의 유사도 계산

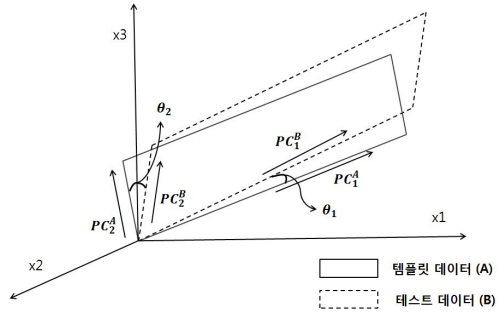
두 초평면이 일치하는 경우	두 초평면이 일치하지 않는 경우
$\cos\theta_{11} = 1$	$\cos\theta_{11} = 0.5$
$\cos\theta_{12} = 0$	$\cos\theta_{12} = 0.5$
$\cos\theta_{21} = 0$	$\cos\theta_{21} = 0.5$
$\cos\theta_{22} = 1$	$\cos\theta_{22} = 0.5$
유사도값: 2	유사도값: 2

그러므로 모든 주성분들 간의 조합에 대해 사이각의 cosine을 계산하지 않고 핵심 주성분들만을 비교하고자 하는데, 앞의 경우에서는 $\cos\theta_{11}$ 과 $\cos\theta_{22}$ 만을 적용하는 것이다.

[표 2] 본 연구에서 제안하는 유사도 계산

두 초평면이 일치하는 경우	두 초평면이 일치하지 않는 경우
$\cos\theta_{11} = 1$	$\cos\theta_{11} = 0.5$
$\cos\theta_{22} = 1$	$\cos\theta_{22} = 0.5$
유사도값: 2	유사도값: 1

[그림 2]는 3차원의 템플릿 데이터 집단과 테스트 데이터 집단의 주성분이 두 개(k=2)인 2차원으로 변환된 경우를 표현한 그림이며, 실선으로 표시한 것이 템플릿 데이터이며 점선으로 표시한 것이 테스트 데이터이다. $\theta_1(\theta_2)$ 은 템플릿 데이터의 첫(두) 번째 주성분 $PC_1^A(PC_2^A)$ 와 테스트 데이터의 첫(두) 번째 주성분 $PC_1^B(PC_2^B)$ 의 사이각이다. 본 연구에서 제안하는 첫 번째 거리척도를 계산하는 방법은 다음과 같다.



[그림 2] 템플릿 데이터와 테스트 데이터의 사이각 비교

$$D1 = \sum_{i=1}^{k} \left| \cos\theta_i^2 - 1 \right| \qquad (2)$$

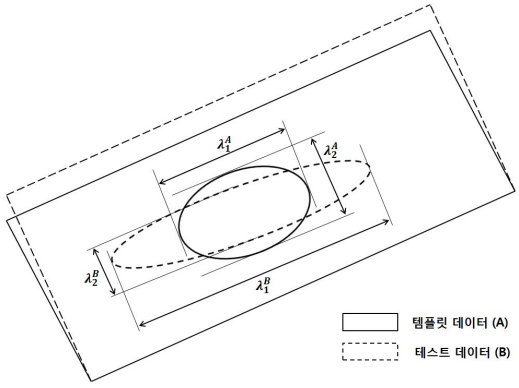
여기서 θ_i 는 템플릿 데이터 집단의 i 번째 주성분과 테스트 데이터 집단의 i 번째 주성분간의 사이각이다. 절대값 안에 -1을 한 이유는 템플릿 데이터 집단과 테스트 데이터 집단과의 주성분 사이각이 0일 때, cosine 값이 1이므로 거리가 0 이 되기 위함이다. 따라서 첫 번째 거리척도 D1의 값은 [0, k] 의 값을 취하고 0에 가까울수록 테스트 데이터는 템플릿 데이터와 유사하다고 말할 수 있다.

2.3 거리척도 2

두 번째 거리척도는 템플릿 데이터 집단의 주성분들로 생성되는 초평면과 테스트 데이터 집단의 주성분들로 생성되는 초평면 위의 score 분포 모양을 비교하는 방법이다. 템플릿 데이터 집단과 테스트 데이터 집단의 두 고유값의 비율을 통해 그 거리를 정의한다.

[그림 3]에서 실선의 타원은 템플릿 데이터 집단의 고유값 λ_1^A 와 λ_2^A 로 표현되고, 점선의 타원은 테스트 데이터 집단의 고유값 λ_1^B 와 λ_2^B 로 표현된다. [그림 3]처럼 만약 주성분들로 생성되는 축 간의 사이각이 유사한 경우, 첫 번째 거리척도로는 정상과 이상을 예측하기 쉽지 않다. 하지만 고유값의 비율로 정의되는 두 번째 거리척도로는 템플릿 데이터 집단과 테스트 데이터 집단의 차이가 존재한다. 두 번째 거리척도는 다음과 같이 정의된다.

$$D2 = \sum_{i=1}^{k} \left| \frac{\min\left(\lambda_i^A, \lambda_i^B\right)}{\max\left(\lambda_i^A, \lambda_i^B\right)} - 1 \right| \quad (3)$$

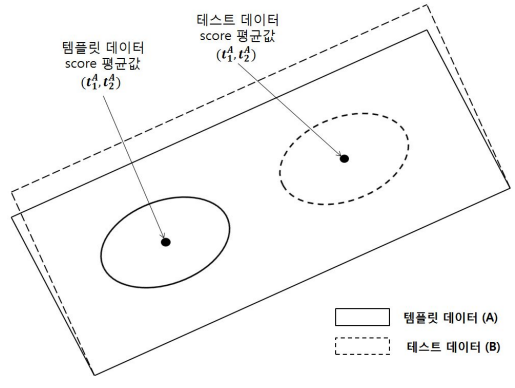


[그림 3] 템플릿 데이터와 테스트 데이터의 score 분포모양 비교

두 번째 거리척도 D2의 λ_i^A 는 템플릿 데이터 집단의 i 번째 고유값이고 λ_i^B 는 테스트 데이터 집단의 i 번째 고유값이다. 템플릿 데이터와 테스트 데이터의 고유값 중 최댓값이 분모에 있고 최솟값이 분자에 있는데 이것은 D2의 범위를 [0, k]로 제한하기 위함이다. D2도 D1과 마찬가지로 0에 가까울수록 테스트 데이터 집단이 템플릿 데이터 집단과 유사하다고 말할 수 있다.

2.4 거리척도 3

두 번째 거리척도가 score 분포 모양의 차이를 나타내었다면, 세 번째 거리척도는 score 분포 위치의 차이를 나타내는 방법이다.



[그림 4] 템플릿 데이터와 테스트 데이터의 score 분포 위치 비교

[그림 4]에서 실선으로 표시한 템플릿 데이터 집단과 점선으로 표시한 테스트 데이터 집단이 유사한 주성분으로 생성된 초평면 위에 같은 score 분포 모양을 띄고 있지만 score의 분포 위치는 다르다. 이 그림에서 템플릿 데이터 집단의 평균값은 $\left(t_1^A, t_2^A\right)$ 이고 테스트 데이터 집단의 평균값은 $\left(t_1^B, t_2^B\right)$ 이라면 두 평균값을 비교한다. 세 번째 거리는 두 평균값의 유클리디안 거리(Euclidean distance)이다.

$$D3 = \sqrt{\sum_{i=1}^{k} \left(t_i^A - t_i^B\right)^2} \quad (4)$$

세 번째 거리척도 D3 의 거리값의 범위는 $[0, \infty]$ 로 0에 가까울수록 템플릿 데이터와 유사하다는 점은 D1 및 D2 와 동일하다.

3. 이상 예측 모델과 경계값 설정

이상을 예측하기 위해 거리척도 D1, D2, D3에 대해 경계값(threshold value)을 설정하고, 그 경계값을 넘어서면 이상으로 간주한다.

3.1 경계값 설정

경계값을 설정하기 위해서 이상치를 찾아내는 Hampel의 이상치(Outlier) 탐지 모델을 이용한다[3][13]. 이 방법은 데이터의 분포를 가정할 필요가 없으며 데이터의 양에 상관없이 경계값을 설정하고 이상을 예측할 수 있다. Hampel 모델을 이용하는 이유는 세 가지 거리 D1, D2, D3가 특정 분포를 따를 것이라고 가정하기 힘들기 때문이다. 각 거리척도에 대해 적용되는 Hampel 모델은 다음과 같다.

① 정상 데이터들의 거리(d_i)을 계산하고 이들의 중앙값($\tilde{d}$) (Median)을 구한다. 여기서 d_i는 i 번째 정상 데이터의 계산된 거리이다.
② 중앙값을 이용하여 정상 데이터 d_i 에 대하여 정상 데이터와 중앙값과의 편차인 r_i 를 구한다.

$$r_i = \left(d_i - \tilde{d}\right) \quad (5)$$

③ $|r_i|$ 의 중앙값 $\widetilde{|r_i|}$ 를 구한다.
④ 경계값 $T = 4.5\widetilde{|r_i|}$ 로 놓는다.
⑤ Hampel 값 $r_i^{'}$ 는 i 번째 테스트 데이터 t_i 와 테스트 데이터의 중앙값 $\tilde{t}$ 과의 편차의 절대값이다.

$$r_i^{'} = \left|t_i - \tilde{t}\right| \quad (6)$$

⑥ $r_i^{'} > T$ 이면 이상으로 예측하고 $r_i^{'} \leq T$ 정상으로 예측한다.

3.2 이상 예측 모델

이상 예측 모델은 크게 두 단계로 진행된다. 우선 데이터 종류에 따른 각 거리척도의 경계값을 설정하고, 경계값을 기준으로 테스트 시계열 데이터 집단의 정상과 이상을 판단하는 것이 이상예측 모델의 진행 순서이다.

거리 척도 D1, D2, D3 의 경계값을 설정하는 일은 다수의 정상 시계열 데이터 집단을 입력받아 템플릿과의 거리를 각각의 척도로 계산한 후 이 결과를 Hampel 모델에 대입하여 결정한다.

테스트 시계열 데이터 집단이 주어졌을 때 정상과 이상을 예측하는 방법은 거리 D1, D2, D3의 Hampel 값을 순차적으로 계산해서 경계값 이하이면 정상이라고 판단하고 아니면 이상이라고 판단한다. 즉, 데이터를 입력받아 거리척도 D1의 Hampel 값을 계산해서 경계값 이상이면 이상으로 판단하고, 경계값 미만이면 거리척도 D2를 이용하여 이상 여부를 다시 따져보고, 이상이 아니라면 거리척도 D3를 이용하여 최종적으로 모두 경계값 미만인 데이터는 정상이라고 판단하고 D1, D2, D3 중 어느 하나라도 Hampel 값이 경계값 이상이면 이상으로 예측하는 방법이다.

4. 실험

4.1 실험 방법

실험 대상은 변수가 세 개이며 각 변수는 100개의 데이터로 시계열을 구성한다. 실험에서 사용하는 시계열 패턴은 세 종류로 선형, 비선형, 변수간 상관 계수가 주어진 경우이다. 각 시계열 종류별로 300개의 정상 데이터를 이용하여 종류별로 D1, D2, D3 의 경계값을 설정한다. 그리고 시계열 종류별로 300 개의 300개의 정상 시계열 데이터 집단과 300개의 이상 시계열 데이터 집단을 생성하여 제안하는 모델의 성능을 시험한다. 정상 데이터 집단은 템플릿 데이터 집단과 유사하며 템플릿 데이터에 약간의 노이즈(Noise)를 주었다. 노이즈는 정규분포를 따르며 평균이 0이고 분산이 1이다. 300개의 이상 시계열 데이터 집단은 모두 정상 데이터와 차이가 있는 데이터로, 100개씩 세 가지 유형의 데이터로 나뉜다.

4.2 실험 결과 및 분석

선형 템플릿을 아래와 같이 정의하였으며, 정상인 300개의 선형 시계열 데이터 집단은 여기에 노이즈를 더하여 생성하였다.

$$x_1 = t$$
$$x_2 = 100 - t$$
$$x_3 = t$$

300개의 이상 시계열 데이터 집단은 100개씩 세 가지의 이상유형이 있다. 이상 유형은 아래의 [표 3]과 같다.

[표 3] 선형 시계열의 이상 유형

	선형		
이상 시계열 데이터 집단 (300개)	유형 1 (100개)	x_1	$x_1 = 2t$
		x_2, x_3	템플릿과 동일
	유형 2 (100개)	x_1, x_3	템플릿과 동일
		x_2	$x_2 = 50 - \frac{t}{2}$
	유형 3 (100개)	x_1, x_2	템플릿과 동일
		x_3	$x_3 = 25\ t \le 50$ $x_3 = 50\ 50 < t \le 100$
정상 시계열 데이터 집단 (300개)		x_1, x_2, x_3	템플릿에 노이즈 포함

선형 시계열에 대한 정상 데이터의 탐지 정확도(Accuracy)는 [표 4]와 같고 여기서 정상 시계열 데이터 집단은 D2를 가지고 100% 정확하게 예측하였다. 이상을 탐지하는 실험에서는 이상 유형 1, 이상 유형 2, 이상 유형 3 모두 D2에서 이상을 100% 정확하게 찾아내었으며 그 결과는 아래 [표 5]와 같다. [표 5]에서 보듯이 D1에서 이상을 탐지 못했다는 것은 비록 이상 시계열 데이터 집단일지라도 정상 시계열 데이터 집단과의 주성분 사이각은 매우 작다는 것을 의미한다. 또한 D3를 가지고도 이상 탐지율이 낮게 나온 것은 분포상 정상과 이상의 평균 위치가 유사하다는 것을 의미하다. 결론적으로 [표 3]의 선형 시계열에서 정상과 이상은 주성분과 평균이 유사하지만 공간상에 퍼져 있는 두 데이터 집단의 모양(고유값)이 매우 다르다는 것을 의미한다.

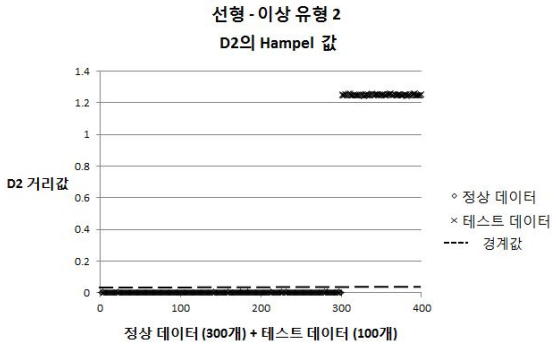
[표 4] 선형 정상 데이터의 탐지 정확도

	선형	정상 탐지율
정상	D1	100%
	D2	98.7%
	D3	100%

[표 5] 선형 이상 데이터의 탐지 정확도

	선형	이상 탐지율	
이상	유형 1 (100개)	D1	0%
		D2	100%
		D3	43%
	유형 2 (100개)	D1	0%
		D2	100%
		D3	0%
	유형 3 (100개)	D1	0%
		D2	100%
		D3	0%

[그림 5]는 선형 정상 시계열 데이터 집단과 이상 유형 2의 시계열 데이터 집단의 거리척도 D2의 Hampel 값을 보여준다. 이 그림에서 점선으로 표시된 선은 경계값이며, 경계값을 초과하는 Hampel 값을 가지면 이상이라고 판단한다.



[그림 5] 선형 정상과 이상 유형 2 데이터의 D2 경계값과 Hampel 값

이번에는 비선형 시계열에 대해서 실험을 진행하였다. 비선형 템플릿을 아래와 같이 정의하였으며, 정상인 300개의 비선형 시계열 데이터 집단은 여기에 노이즈를 더하여 생성하였다.

$$x_1 = \sqrt{t}$$
$$x_2 = 5\sin(t) + 5$$
$$x_3 = 10 - 5\log(t)$$

300개의 이상 시계열 데이터 집단은 [표 6]과 같이 세 가지의 이상 유형을 100개씩 만들어 실험하였다.

[표 6] 비선형 시계열의 이상 유형

	비선형		
이상 시계열 데이터 집단 (300개)	유형 1 (100개)	x_1	$x_1 = 2\sqrt{t}$
		x_2, x_3	템플릿과 동일
	유형 2 (100개)	x_1, x_3	템플릿과 동일
		x_2	x_2 의 노이즈가 N(0,25)
	유형 3 (100개)	x_1, x_2	템플릿과 동일
		x_3	템플릿과 주성분은 동일하지만 고유값은 다른 데이터
정상 시계열 데이터 집단 (300개)		x_1, x_2, x_3	템플릿에 노이즈 포함

[표 7]에 보듯이 비선형 정상 시계열 데이터 집단은 D2에서 100% 정확하게 정상을 예측하였다. 이상을 탐지하는 실험에서는, [표 8]에서 보듯이, 이상유형 1은 D1 과 D2, 이상유형 2와 이상 유형 3은 D2에서 이상을 정확하게 찾아내었다. 따라서 이상 유형 1의 시계열 데이터 집단은 정상 집단과 비교해 봤을 때 주성분의 사이각뿐만 아니라 분포의 모양(고유값)도 매우 다르다는 것을 의미한다. 그리고 이상 유형 2와 3의 경우는 주성분의 사이각과 평균의 위치는 유사하지만 분포의 모양이 다르다는 것을 알 수 있었다.

[표 7] 비선형 정상 데이터의 탐지 정확도

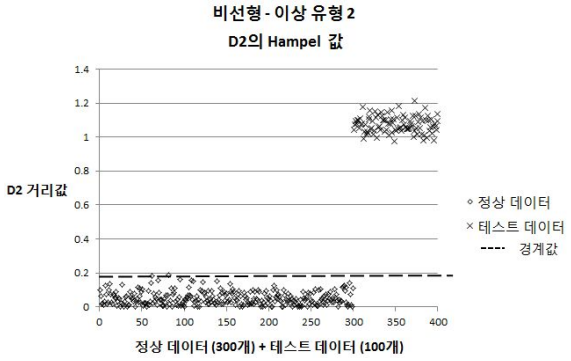
	비선형	정상 탐지율
정상	D1	90%
	D2	100%
	D3	90%

[표 8] 비선형 이상 데이터의 탐지 정확도

	비선형	이상 탐지율	
이상	유형 1 (100개)	D1	100%
		D2	100%
		D3	0%
	유형 2 (100개)	D1	12%
		D2	100%
		D3	0%
	유형 3 (100개)	D1	19%
		D2	100%
		D3	0%

[그림 6]은 비선형 정상 시계열 데이터 집단과 이상 유형 2의 시계열 데이터 집단의 거리척도 D2의 Hampel 값을 보여준다. 여기서 이상 데이터 집단은 모두 경계값 위에 위치한

것을 확인할 수 있다.



[그림 6] 비선형 정상과 이상 유형 2 데이터의 D2 경계값과 Hampel 값

마지막으로 변수간에 상관관계가 존재하는 시계열 데이터를 대상으로 이상 탐지 실험을 진행하였다. 상관관계를 갖는 데이터의 템플릿을 아래와 같이 정의하였고, 정상인 300개의 데이터는 여기에 노이즈를 더하여 생성하였다.

$$Corr(x_1, x_2) = -0.9 \quad \text{(음의 상관관계)}$$
$$Corr(x_1, x_3) = 0.7 \quad \text{(양의 상관관계)}$$

300개의 이상 시계열 데이터 집단은 [표 9]와 같이 100개씩 세 가지의 이상 유형을 생성하였다.

[표 9] 상관관계를 가지는 시계열의 이상 유형

	상관 관계	
이상 시계열 데이터 집단 (300개)	유형 1 (100개)	$Corr(x_1, x_2) = -0.1$ $Corr(x_1, x_3) = 0.7$
	유형 2 (100개)	$Corr(x_1, x_2) = -0.9$ $Corr(x_1, x_3) = -0.7$
	유형 3 (100개)	$Corr(x_1, x_2) = -0.9$ $Corr(x_1, x_3) = 0.1$
정상 시계열 데이터 집단 (300개)	템플릿에 노이즈 포함	

상관관계를 가지는 시계열 데이터 집단의 정상 탐지율은 [표 10]와 같으며 D1과 D2를 가지고 99.4%의 탐지 정확도를 보였다. 이상을 예측하는 실험에서는 [표 11]에서 보듯이 이상 유형 1, 이상 유형 2, 이상 유형 3 모두 D1에서 100% 이상을 정확하게 예측하였다.

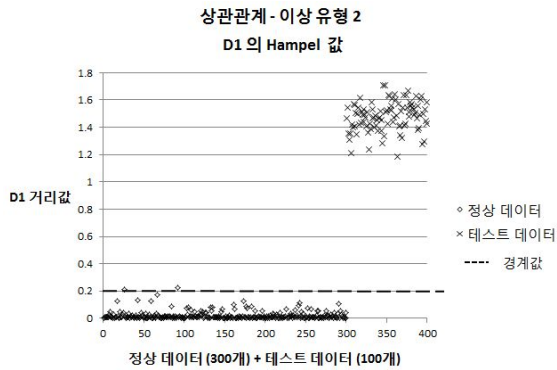
[표 10] 상관관계를 갖는 정상 데이터의 탐지 정확도

	상관관계	정상 탐지율
정상	D1	99.4%
	D2	99.4%
	D3	82%

[표 11] 상관관계를 갖는 이상 데이터의 탐지 정확도

	상관관계	이상 탐지율	
이상	유형 1 (100개)	D1	100%
		D2	89%
		D3	100%
	유형 2 (100개)	D1	100%
		D2	100%
		D3	100%
	유형 3 (100개)	D1	100%
		D2	76%
		D3	100%

[그림 7]은 정상 상관관계를 갖는 시계열 데이터 집단과 이상 유형 2의 시계열 데이터 집단의 거리척도 D1의 Hampel 값을 보여준다.



[그림 7] 상관관계를 가지는 정상과 이상 유형2 데이터의 D1 경계값과 Hampel 값

세 가지 시계열 패턴 종류를 대상으로 실험을 한 결과 거리 척도 D1, D2, D3를 이용하면 모두 정상과 이상을 99%이상 정확하게 탐지할 수 있음을 알았다. 이 결과로부터 본 연구에서 제안한 거리 척도의 우수성을 확인하였다.

5. 결론 및 추후 연구 방향

본 연구에서는 템플릿 시계열 데이터 집단과 테스트 시계열 데이터 집단이 주어졌을 때 각 데이터 군을 주성분 분석을 통해서 주성분 공간내의 score 분포로 표현하고 두 분포의 특징(주성분간의 사이각, 고유값의 비율,

score 평균값의 비율)을 비교해서 거리 척도 D1, D2, D3로 표현하고 Hampel 값을 이용하여 이상 여부를 탐지하는 방법을 제안하였다. 선형관계, 비선형 관계, 임의의 상관관계가 있는 시계열 데이터 집단에 대해 제안한 거리 척도를 적용한 결과 이상 여부를 매우 정확히 예측하였다.

본 연구에서 제안한 주성분 분석은 데이터의 정규분포 가정이 요구되지 않으며 이상 여부를 결정하는 경계값 결정에도 분포와 무관한 Hampel 방법을 적용하였기 때문에 활용도가 높다고 할 수 있다. 하지만 본 연구에서 사용한 데이터는 실험에서 밝혔듯이 임의로 만든 데이터라는 한계를 갖는다. 따라서 반도체 생산 공정과 같이 실제 상황에서 얻어진 데이터를 적용하여 제안한 방법의 성능을 평가할 필요가 있다. 그리고 향후 연구에서는 다음 두 가지 측면의 개선이 필요하다. 첫째, 실험에 사용한 데이터는 시계열의 길이와 변수의 범위가 모두 동일하다고 가정하였다. 그러나 실제 상황에서는 시계열의 길이와 변수의 범위가 다르기 때문에 정규화하는 전처리 과정이 필요하다. 둘째, 실제 상황에서는 시계열을 구성하는 변수마다 이상에 영향을 미치는 정도가 다르다. 따라서 변수 선택과정을 거친 후 선택된 변수들의 시계열 데이터 집단을 가지고 세 가지 거리 척도를 계산하는 과정이 필요할 것으로 예상된다.

사사

이 논문은 2012년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2012R1A1A2046061).

참고문헌

[1] Baek, J. G., Kim, C. O., Kim, S. S. (2002) Online learning of the cause-and-effect knowledge of a manufacturing process, *International Journal of Production Research*, 40(14), 3275-3290.

[2] Barna, G. G. (1992) Automatic problem detection and documentation in a plasma etch reactor, *IEEE Transactions on Semiconductor Manufacturing*, 5(1), 56-59.

[3] Ben-Gal, I. (2005) *Outlier detection, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic Publishers.

[4] Goodlin, B. E., Bonig, D. S., Sawin, H. H., Wise, B. M. (2003) Simultaneous fault detection and classification for semiconductor manufacturing tools, *Journal of the Electrochemical Society*, 150(12), 778-784.

[5] Guo, H.-F., Spanos, C. J., Miller, A. J. (1991) Real time statistical process control for plasma etching, *IEEE/SEMI International*, 113-118.

[6] He, Q. P. and Wang, J. (2007) Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes, *IEEE Transactions on Semiconductor Manufacturing* 20(4), 345-354.

[7] Kim, B. and May, G. S. (1997) Real-time diagnosis of semiconductor manufacturing equipment using neural networks, *IEEE Transactions on Components, Packaging and Manufacturing Technology-Part C*, 20(1), 39-47.

[8] Ko, J. M., Hong, S. R., Choi, J. Y., Kim, C. O. (2013) Wafer-to-wafer process fault detection using data stream mining techniques. *International Journal of Precision Engineering and Manufacturing*, 14(1), 103-113.

[9] Ko, J. M., Kim, C. O., Lee, S. J., Hong, J. P (2010) Structural feature-based fault-detection approach for the recipes of similar products, *IEEE Transactions on Semiconductor Manufacturing,* 23(2), 273-283.

[10] Ko, J. M. and Kim, C. O. (2012) A multivariate parameter trace analysis for online fault detection in a semiconductor etch tool, *International Journal of Production Research*, 50(23), 6639-6654.

[11] Krzanowski, W. J (1979) Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74(367), 703-707.

[12] Kun, Lin. (2012) A Multi-dimensional time series data mining model for weather forecast. *Advanced Materials Research*, 532-533, 1277-1281.

[13] Redman, T. C. (2001) *Data Quality: The Field Guide*, Boston Digital Press.

[14] Sarmiento, T., Hong, S. J., May, G.S. (2005) Fault detection in reactive ion etching systems using one-class support vector machines, *Advanced Semiconductor Manufacturing Conference and Workshop, IEEE/SEMI*, 139-142.

[15] Son, I. S. and Oh, K. J. (2009) An early warning system for global institutional investors at emerging stock markets based on machine learning forecasting, *Expert Systems with Applications,* 36(3), 4951-4957.