ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 45, No.1:025-038, March 2017
http://dx.doi.org/10.7469/JKSQM.2017.45.1.025

군집기반 열간조압연설비 상태모니터링과 진단

서명교* · 윤원영**†

*포스코
**부산대학교 산업공학과

Clustering-based Monitoring and Fault detection in Hot Strip Roughing Mill

MYUNG-KYO SEO* · WON YOUNG YUN**†

*POSCO
**Department of Industrial Engineering, Pusan National University

ABSTRACT

Purpose: Hot strip rolling mill consists of a lot of mechanical and electrical units. In condition monitoring and diagnosis phase, various units could be failed with unknown reasons. In this study, we propose an effective method to detect early the units with abnormal status to minimize system downtime.

Methods: The early warning problem with various units is defined. K-means and PAM algorithm with Euclidean and Manhattan distances were performed to detect the abnormal status. In addition, an performance of the proposed algorithm is investigated by field data analysis.

Results: PAM with Manhattan distance(PAM_ManD) showed better results than K-means algorithm with Euclidean distance(K-means_ED). In addition, we could know from multivariate field data analysis that the system reliability of hot strip rolling mill can be increased by detecting early abnormal status.

Conclusion: In this paper, clustering-based monitoring and fault detection algorithm using Manhattan distance is proposed. Experiments are performed to study the benefit of the PAM with Manhattan distance against the K-means with Euclidean distance.

Key Words: Hot strip rolling mill, Fault detection and classification, K-means, PAM

• Received 9 January 2017, 1st revised 22 January 2017, accepted 23 January 2017
† Corresponding Author(wonyun@pusan.ac.kr)

1. 서 론

철강산업에서 제품의 과잉공급과 치열한 시장경쟁으로 인해 원가절감은 매우 중요한 문제로 대두되고 있다. 이에 따라 원가경쟁력을 확보하기 위한 설비보전 비용절감과 가동률 증대는 필수적이다. 이를 위해서는 설비의 상태를 미리 파악하고 경제적으로 설비를 보전하는 연구가 필요하다. 본 연구에서는 철강제조공정에서 수집되는 데이터를 이용하여 설비상태의 건전성 여부(healthy monitoring)를 파악할 수 있는 군집기반 설비이상 분석 절차를 제안함으로써 비계획 설비가동중단(unscheduled shutdown)에 의한 가동률 하락과 경제적 손실의 최소화를 목적으로 한다.

본 논문에서 대상으로 하는 설비인 열간조압연기(hot strip roughing mill)는 재결정온도(recrystalizing temperature)이상으로 가열된 슬라브(slab)를 가공하는 설비이다. 조압연공정은 두껍고 폭이 큰 슬라브를 압연하므로 다른 공정에 비해 상대적으로 큰 가공부하를 가진다. 조압연에서 가공된 중간재(bar)는 마무리압연(hot strip finishing mill)의 통판성과 품질을 좌우하며 bar형상이 권취(coiling)성능에도 영향을 미친다. 또한 조압연설비가 비계획적으로 정지되는 경우 가열로에서 추출된 슬라브는 대기중에서 냉각되므로 재열처리(reheating)해야 하며 가열로 내부 슬라브 또한 적당한 시점에 인출되지 못하게 되어 품질 불균일의 원인이 된다. 따라서 조압연설비의 이상여부를 사전에 정확히 탐지하는 것은 불필요한 비계획 보전비용을 줄이고 품질과 생산성을 높이는데 중요한 문제이다.

이상진단(fault detection)의 목적은 반응변수 혹은 출력변수(output)에 대한 예측력을 최대로 하는 목적함수 예측방법을 선택 또는 개발하는 것이다. '설비가 정상 상태에서 얼마나 벗어나 있는가', '설비가 이전상태와는 다른 상태인가'에 대한 문제는 데이터마이닝 관점에서 보면 시계열 데이터로 구성된 설비변수(equipment parameter)를 독립변수 X로 하는 이진분류(binary classification)문제, 지도학습(supervised learning)문제로 정의할 수 있다. 하지만 설비의 정상, 이상상태 값, 즉, 반응변수 Y가 주어지지 않는 경우에는 분류모델의 지도 학습데이터로 사용할 수 없다. 본 논문에서는 비지도학습(unsupervised learning)문제인 군집기반 모델을 통해 정상상태에서 나타나지 않던 이상군집 검출을 통해 반응변수 Y를 설비이상유무로 분류(classification) 할 수 있음을 보인다. 여기서 반응변수 Y는 클래스 값(정상, 비정상)을 가지는 이진변수이다.

시계열 데이터형태를 띠는 설비상태감시 변수는 ETL(extraction, transformation, loading)이라는 데이터 수집, 변환, 저장장치에 의해 데이터베이스에 실시간으로 저장된다. ETL장치에서 수집되는 설비변수의 종류가 많을 뿐만 아니라 ms(millisecond)주기로 수집되므로 저장되는 데이터의 양이 매우 크다. 이러한 대용량의 데이터를 효율적으로 이용하기 위해 데이터를 일정 시간 간격으로 나누고 이를 평균, 표준편차, 왜도(skewness), 첨도(kurtosis) 등의 변수로 파생(derived parameter)시켜 데이터를 축약할 필요가 있다. 본 논문에서 제안하는 군집기반 설비이상 탐지 알고리즘(clustering-based fault detection algorithm)절차는 다음과 같은 특성을 갖는다.

첫째, 군집기반 설비이상 탐지절차를 제안한다. 군집기반 분석방법은 구현이 비교적 용이하고 방법론에 대한 직관적인 이해와 설명이 용이하다.

둘째, 비계층적 군집분석(non-hierarchical clustering analysis)알고리즘 중 K-means와 PAM (partition around medoids) 알고리즘을 이용해서 설비이상을 탐지하는 절차를 제안한다. K-means와 PAM은 다음과 같은 특징을 갖는다.

(1) K-means는 비계층적 군집 알고리즘 가운데 간단하면서도 구현이 용이하다. PAM은 K-means에 비해 알고리즘은 다소 복잡하지만 축약된 파생 변수를 이용할 경우 적용이 가능하다.

(2) K-means는 군집의 중심좌표(centroids)를 사용하며, 간단한 거리 계산을 통하여 군집화가 가능한 반면 이상치(outlier)와 잡음(noise)에 민감한 단점을 가진다. 반면 PAM은 군집의 중심개체(medoids)를 사용하므로

K-means에 비해 이상치와 잡음에 둔감(robust)한 장점을 가진다.

(3) Hotelling T^2 관리도와 판별분석(discriminant analysis)의 경우 분포의 형태를 가정하지만 K-means와 PAM 알고리즘은 어떤 형태의 분포에 대한 가정도 필요로 하지 않는다.

셋째, 최적 군집수 결정에 관한 다양한 지표를 이용할 경우 지표마다 상이한 결과를 얻을 수 있다. 본 논문에서는 다양한 군집수 결정 지표에서 선정된 최적 군집수를 다수결의 원칙(majority rule)에 따라 선정하였다.

넷째, 군집분석 결과를 전체 군집성능과 개별 군집의 성능을 알아보기 위해 실루엣 폭(silhouette width)과 오분류 데이터 갯수로 정량화하여 군집성능을 평가하였다.

군집분석을 통한 설비 이상탐지 문제는 전자제품, LCD, 반도체, 항공기엔진, 풍력발전기 등 다양한 산업에서 연구, 적용되어왔다. 대표적인 관련 사례연구들을 보면 Kang and Kim(2013)의 연구에서는 군집분석(clustering analysis)을 기반 한 관리도를 이용하여 설비이상을 탐지했다. 이 연구에서는 모니터링 통계량으로서 거리척도를 이용하여 새로운 관측값의 이상(abnormality)의 정도가 기존 군집으로부터의 최소거리를 가지는 군집에 할당하였다. Hotelling T^2 관리도의 한계를 지적후 거리에 기반한 설비이상 모니터링 방법을 제안하였다. 또한 알고리즘에 대한 유효성을 TFT-LCD제조공정에 적용하여 검증하였다. Verdier and Ferreira(2011)와 Kwak and Kim(2013)는 반도체 공정의 이상을 검출하기 위해 적응형 거리에 기반으로 k-nearest neighbor rule을 제안하였으며 Kumar et al.(2010)은 전자기기의 고장 탐지를 위해 오차함수를 최소화하는 한계값(threshold value)을 이용하여 건전성 여부를 판단하였다. Bharambe et al.(2014)는 비지도 기계학습 알고리즘인 K-means와 주성분분석(principal component analysis, PCA)의 차원축소 기법을 접목하여 이상치를 검출하는 방법을 제안하였다. 다만, 거리의 척도를 사용하였으나 K-means 알고리즘 내에 사용한 것은 아니고 이상치 검출에만 사용하였다.

본 연구에서는 설비가 운전중에 정상상태로부터 이탈했는지 여부를 탐지하는 군집기반 PAM 알고리즘 적용 절차를 제안하고자 한다. 본 논문의 나머지 부분은 다음과 같이 구성된다. 2장에서는 군집기반 설비이상 탐지절차를 소개하고, 3장에서는 실험 및 결과고찰을 통해 제안된 절차의 유효성을 검증하고 결과를 해석한다. 마지막으로 4장에서는 연구결과를 정리하고 제안된 절차의 의의를 논의한다.

2. 군집기반 설비이상 탐지절차

2.1 군집화 알고리즘

군집화(clustering)란 사전지식 없이 관측값 또는 개체를 의미 있는 몇 개의 부분집단으로 나누는 과정을 말한다. 여기서 나뉜 부분집합을 군집(cluster)이라고 하며 군집화에서 의미 있는 군집이란 같은 집단에 속한 관측값 또는 개체들이 서로 유사하고 다른 군집에 속한 개체 사이에는 유사성이 적은 것을 의미한다. 전체 자료를 군집화하면 각 개체에 대한 상세함을 잃어버리게 되지만 전체를 간단히 표현 할 수 있는 장점이 있다. 일반적으로 군집화가 잘 되었는지를 평가할 수 있는 변수가 없기 때문에 군집분석은 데이터마이닝에서는 자율학습 알고리즘으로 분류된다. 비계층적 군집방법의 대표적인 방법으로 McQueen(1967)과 Hartigan and Wong(1979) 등이 제안한 K-means 군집화가 있다. 최근 데이터의 대용량화로 인해 계층적 군집화나 그래프이론(graph theory)을 이용한 군집화로는 처리시간 복잡도 측면에서 비효율적이다. 본 연구에서는 K-means 및 PAM 알고리즘을 기반으로 설비이상 패턴을 검출하기로 한다.

2.2 비계층적 군집화알고리즘

2.2.1 K-means 알고리즘

K-means 알고리즘은 구현이 쉽고 속도가 빠르므로 가장 많이 사용되는 비계층적 군집분석 방법이다. K-means의 기본 개념은 패턴들과 그 패턴이 속하는 클러스터의 중심과의 평균거리를 최소화하는 것이다. 즉, K-means 알고리즘은 주어진 데이터를 k 개의 군집으로 묶는데 있어, 각 군집과 거리 차이의 분산을 최소화하는 방식으로 동작한다. 이 알고리즘은 자율학습의 일종으로 레이블이 달려 있지 않은 입력데이터에 레이블을 달아주는 역할을 수행한다. 그룹을 나누는 과정은 거리 기반의 그룹간 비유사도(dissimilarity)를 비용 함수(cost function)로 이용하며 이를 최소화하는 방식으로 이루어지며 이 과정에서 같은 그룹 내 데이터 개체끼리의 유사도는 증가하고, 다른 그룹에 있는 데이터 개체와의 유사도는 감소하게 된다. K-means 군집화는 데이터의 수(n)가 늘어나면 적절한 시간 내에 최적해(optimal solution)를 찾는 것은 불가능하여 근사해를 제공하는 알고리즘이다. 일반적인 유클리드 공간에서 K-means 문제의 최적해를 찾는 것은 NP-hard 문제로 알려져 있다. 이상 언급한 K-means 알고리즘의 특징은 다음과 같다.

1. 군집수 k를 입력파라미터로 지정해 주어야하며 k에 따라 군집성능이 크게 달라진다.
2. 초기값에 따라 최적화결과가 전역 최적해가 아닌 지역 최적해에 빠질 가능성이 있다.
3. 이상치(outlier)에 민감하다. 이상치는 알고리즘 내에서 중심점을 갱신하는 과정에서 군집 내의 전체 평균값을 크게 왜곡시킬 수 있다.
4. 구형(spherical)이 아닌 군집을 찾는 데에는 적절하지 않다.

2.2.2 PAM 알고리즘

K-means 알고리즘에서는 각 군집의 중심좌표(centroid)를 고려하고 있는 반면, K-medoids 군집방법의 한 종류인 PAM에서는 각 군집의 대표개체(medoid)를 고려하는 기법이다. 중심좌표를 사용하는 것보다 대표객체를 사용하는 것이 이상치에 덜 민감한 장점을 가진다. 군집의 대표개체란 그 군집에 속하는 개체 중 다른 개체들과의 평균(또는 전체) 거리가 최소가 되는 개체를 말한다. K-medoids 방법 역시 K-means와 마찬가지로 적절한 시간내에 최적해를 구하는 것은 어려운 것으로 알려져 있다. 따라서 K-medoids 알고리즘 중 하나인 PAM과 같은 발견적 해법을 사용한다. K-medoids방법에서 대표적인 PAM 알고리즘은 이상치와 잡음에 대해 둔감한 특성을 보이는 것 외에 K-means알고리즘의 특징을 모두 가진다.

Table 1. K-means algorithm

K-means algorithm
Step 1 : **(Initial object selection)** By some rule, the coordinates of k objects are selected as the centroid of the initial cluster.
Step 2 : **(Cluster assignment of objects)** For each object, calculate the distance to the center of the cluster and then assign the object to the closest cluster.
Step 3 : **(Calculation of cluster center coordinates)** Calculation of cluster center. coordinates.
Step 4 : **(Convergence condition check)** The newly calculated center coordinate value is compared with the previous coordinate value to be within the convergence condition, otherwise step 1 is repeated.

2.3 최적 군집수(k)의 결정

군집수 k를 결정하는 많은 지표들이 개발되어 있다. 본 논문에서는 대용량데이터분석에 사용되는 오픈소스웨어 프로그램인 R에서 제공하는 패키지, NbClust(Charrad et al. 2014)를 이용하여 최적 군집개수를 선정하고자 한다. 또한 Gap통계량을 그래프에 나타내어 군집 개수 k를 직관적으로 확인 할 수 있도록 한다. Gap통계량은 식[1]와 같이 정의된다.

$$Gap_n(k) = E_n^* \log(W_k) - \log(W_k) \quad (1)$$

여기서, E^*는 표본크기 n인 기준분포로부터 얻어진 기댓값을 의미하고, W_k는 오차의 척도로써 군집내 제곱합(within-cluster sum of square)으로 정의되며 군집내 개체가 얼마나 흩어져 있는지를 의미한다. Kaufman and Rousseeuw(1990)는 실루엣 크기, 식[2]와 같이 $s(i)$를 정의하고 $s(i)$를 최대화하는 k값을 최적 군집갯수로 제안하였다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

여기서, $a(i)$는 개체 i로부터 동일한 군집내의 모든 다른 개체들과의 평균이며, $b(i)$는 개체 i와 다른 군집내 모든 다른 개체들과의 평균거리이다. 이때 계산된 $s(i)$는 $-1 \le s(i) \le 1$ 의 값을 가지며 1에 가까울수록 i는 올바른 군집에 분류된 것이며 -1에 가까울수록 잘못된 군집에 분류되었음을 나타낸다. 본 논문에서는 Gap통계량을 사용하여 최적 군집갯수 k를 결정하고 선정된 군집수에서 관측되는 설비변동을 모니터링한다.

Table 2. The main characteristics of the K-means and PAM

Methods	Type	Complexity[a]	Input	Result	Cluster Criterion
K-means	Numerical	$O(n)$	k	Centerid	$\min_{v_1, v_2, \ldots, v_k}(E_k), E_k = \sum_{i=1}^{k}\sum_{k=1}^{n} d^2(x_k, v_i)$
K-Medoids (PAM)	Numerical	$O(k(n-k)^2)$	k	Medoid	$\min(TC_{ih}),\ TC_{ih} = \sum_j (C_{jih})$

[a] n is the number of points in the dataset and k the number of clusters defined.

2.4 거리척도(distance measure)의 선정

군집분석의 기본 아이디어는 한 군집 내의 개체들 간의 유사성은 가능한 크게, 서로 다른 군집간의 유사성은 가능한 작도록 군집들을 형성하는 것이다. 본 논문에서는 유클리드거리와 맨하탄거리를 거리척도로 사용하여 제안된 알고리즘의 성능을 비교하고자 한다.

Table 3. PAM algorithm

PAM Algorithm
[BUILD]
Step 1 : The distances between different individuals are obtained for each individual, and a single object having the smallest sum is selected as a medoid. Let M be the selected representative entity set.
Step 2 : For entity j not selected as representative entity, find the distance D_j closest to entity j among the individuals previously selected as representative entity. That is, $D_j = \min_{k \in M} d(j,k), \quad j \notin M$ Then, for two individuals i and j not selected as representative individuals, the following is calculated. $C_{ij} = \max(D_j - d(j,i), 0) \quad i,j \notin M$
Step 3 : Include the object m with the largest distance reduction as follows in the representative object and modify the representative object set.
Step 4 : If k representative objects have been selected, move to SWAP step, otherwise return to step 1.
[SWAP]
Step 1 : In order to calculate the change of objective function when exchanging entity i and entity h, we first calculate the change in any entity j ($j \neq h$) that is not selected as a representative entity as follows. C_{jih} =(After exchanging i and h, distance between object j and representative object) - (The distance between object j before the exchange and the representative object) $(j \notin M, i \in M, h \notin M)$
Step 2 : When the representative entity i is exchanged for h, the total variation is as follows. $T_{ih} = \sum_j C_{jih}$ At this time, if object i^* and j^* corresponding to $\min_{i,h} T_{ih}$ are found, $T_{i^*h^*} < 0$ returns to step 1 after exchange and if $T_{i^*h^*} \geq 0$ is terminated without exchange.

3. 실험 및 결과고찰

철강제품에서 열연코일은 중요한 최종 혹은 중간제품으로 사용된다. 중간제품으로 사용되는 경우에는 열연코일을 가공하여 냉연, 전기강판, 도금, STS냉연 강판 등 최종제품이 생산된다. 또한 열연제품의 가공품질은 최종제품의 통판성과 품질을 좌우하게 되므로 공정에서의 중요성이 크다. 열연공정은 가열(heating), 조압연(roughing mill), 마무리압연(finishing mill), 냉각(cooling) 및 권취(coiling)공정 등 크게 4개 공정으로 대별된다. 특히 조압연공정은 두꺼운 슬라브를 마무리압연이 가능한 두께와 폭으로 가공하므로 상대적으로 큰 부하를 받는 설비이다. 조압연설비의 이상은 소재가 좌우로 휘게 되는 캠버(camber), 상하로 휘게 되는 상향 혹은 하향을 발생시켜 후속공정인 사상압연공정의 성능에 직접적인 영향을 주므로 설비 이상유무를 조기에 판단하는 것이 중요하다. 조압연 설비는 압연롤과 동력을 전달하는 축, 커플링, 동력을 생산하는 모터로 구성된다. 여기서 수집되는 전류, 토크, 온도 등의 데이터를 통해 설비의 가동상태가 정상인지 여부를 모니터링하고 이상시 빠르게 검출해내는 것이 중요하다. 본 연구에서는 조압연 설비의 상태파악과 관련하여 수집된 센서데이터를 가공하여 정상상태와 이상상태를 구분하고자 한다. 이를 통해 설

비의 계획되지 않은 중단(unscheduled breakdown)을 미연에 방지하고 필요할 경우 예지보전을 실시하여 비계획 설비중단으로 인한 손실을 최소화 하는 것을 목적으로 한다.

3.1 군집기반 설비모니터링 및 진단절차

군집분석기반으로 설비상태를 모니터링하고 진단하기 위해서는 먼저 수집된 데이터를 전처리하고 무부하구간 데이터를 추출해야한다. 본 논문에서는 군집기반 설비상태모니터링과 진단절차를 Figure 1과 같이 제시한다. 데이터 분석을 위해서는 수집된 데이터의 전처리가 선행되어야한다. 그리고 설비가 부하가동상태(load operation)에서는 압연되는 재료의 물성에 따라 측정데이터가 변동하므로 설비가 다음 압연을 위해 무부하가동(no load operation) 시간대의 데이터만을 별도 추출한다. 데이터 변동에 대한 깊이 있는 분석을 위해 평균, 표준편차, 왜도, 첨도 값을 일정주기를 가지고 계산한다. 이들 데이터의 특징과 구조에 대해 탐색하고 통찰을 얻기 위해 탐색적 데이터분석 즉, EDA(exploratory data analysis)를 수행한다. 이를 통해 데이터의 개략적인 분포의 경향과 이상치 유무, 상관관계 등을 파악한다. 비계층적 군집분석(non-hierarchical cluster analysis)을 실시하기 위해서는 적합한 군집갯수를 지정하는 것이 군집성능에 큰 영향을 미친다. 따라서 데이터 집합을 가장 잘 구분하는 최적 군집갯수 k를 결정해야 한다. 선정된 군집 갯수를 입력값으로 군집분석을 수행후 군집분석의 성능을 평가한다. 분석의 타당성과 적합성을 평가하기 위해 확증적 데이터 분석인 CDA(confirmatory data analysis)를 수행하게 되는데 그래프분석과 통계량 분석을 통해 군집분석 알고리즘에 따른 군집성능을 비교하게 된다. 마지막으로 군집분석 결과를 해석하고 설비성능의 이상상태여부를 판단한다.

실제 열간조압연 설비에서 시간에 따라 수집된 5개의 측정변수를 분석에 사용하였다. 각 측정변수마다 50개의 데이터를 묶어 평균, 표준편차, 왜도, 첨도변수를 파생시킨 결과 20개 변수에 대한 232개 데이터 셋을 구성하였다.

공정이상을 탐지하는 통계적 방법으로 제안된 다변량분석 방법을 이용하여 분석후 제안된 방법과 성능비교를 실시하고자 한다. 다변량분석 방법에서 대표되는 Hotelling T^2관리도에서는 여러 변수를 하나의 지표로 나타내고 이를 관리도에 도시함으로써 전체 데이터가 관리범위를 벗어나는지 파악할 수 있다. 본 연구에서는 먼저 Hotelling에 제안한 다변량 통계적공정관리도(multivariate statistical process control chart, MSPC)를 통해 설비이상이 탐지되는지 관찰하였다. Hotelling T^2 관리도의 경우 데이터가 다변량 정규분포를 따르는 가정이 있어야만 좋은 성능을 보장한다. Figure 2은 고려된 모든 변수를 대상으로 관리도를 도시한 결과를 나타낸다. Figure 2의 좌측 그래프에서 보는 바와 같이 설비이상이 발생한 시점에 일시적으로 관리상한(upper control limit, UCL)를 벗어난 직후 관리한계내로 복귀하는 것을 볼 수 있다. 정규성 검정결과의 일부를 Table 5에 나타내었으며 이 중 정규분포를 따르는 변수만을 대상으로 Hotelling T^2 관리도를 Figure 2의 우측에 도시하였다. Figure 2의 오른쪽 그래프를 보면 설비 이상이 발생한 시점에 관리한계를 벗어나고 이후에도 불안정한 변동을 보이면서 관리한계를 종종 벗어나는 것을 볼 수 있다. 그러나 일관되게 이상신호를 나타내지 못하는 점에서는 모든 변수를 고려한 관리도와 크게 다르지 않음을 알 수 있다.

수집된 변수에 대한 데이터의 형태를 확인하기 위해 Figure 3에서 히스토그램과 Q-Q플롯 결과를 도시하였으며 이상데이터가 분포의 꼬리부분에 많이 나타나는 것을 확인할 수 있다. Pearson's Chi-square 및 Anderson-Darling검정결과 X11, X41은 정규분포 가정이 만족하지만 나머지 변수는 정규분포를 따르지 않음을 알 수 있다. 대부분의 변수가 정규분포를 따르지 않으므로 정규분포를 가정하는 다변량공정관리의 대표적 방법인 Hotelling T^2 관리도로는 설비이상을 적절하게 모니터링 할 수 없음을 알 수 있다. Figure 4는 평균으로 파생된 변수에 대한 산포도, 히스토그램, 상관계수를 나타내고 있다. X21과 X31이 0.67로 비교적 높은 상관관계를 나타내었다.

또한, 산점도에서 설비정상(solid dot)과 이상 데이터(empty dot)가 분포를 달리하고 있음을 파악할 수 있다. EDA를 실시한 후 최적 군집수 결정, 군집분석 및 성능평가, 결과해석을 실시한다.

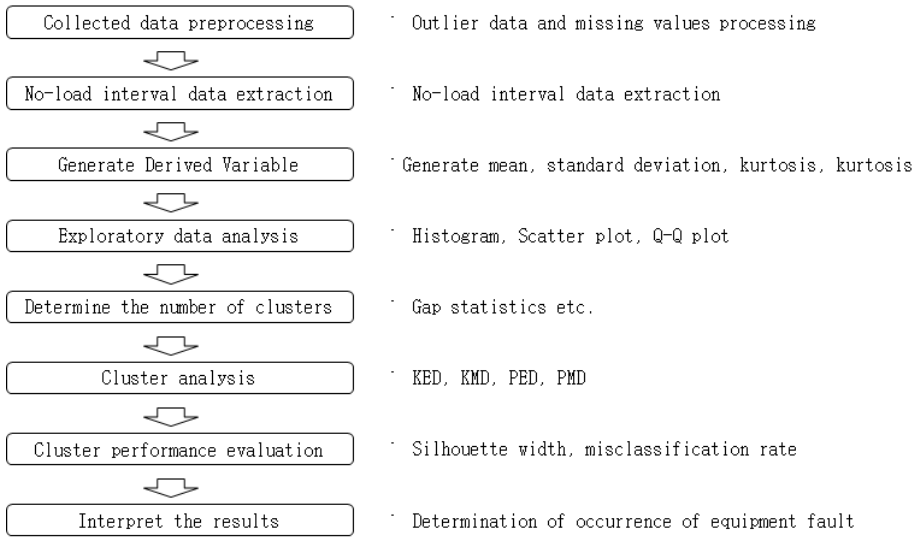

Figure 1. Cluster analysis based facility status monitoring and diagnostic procedures

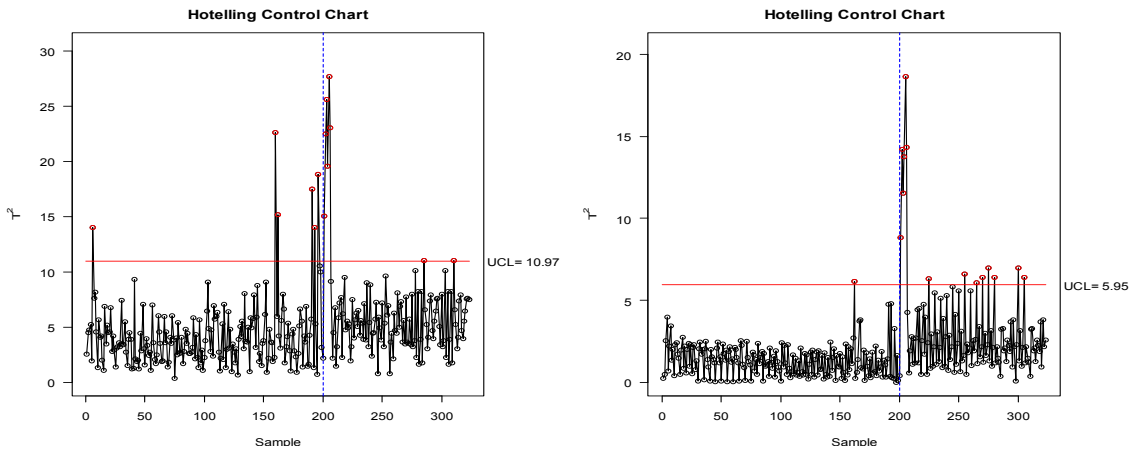

Figure 2. Hotelling control chart(left: All variables, right: variables with normal distribution)

Table 4. Normality test by Pearson chi-square and Anderson-Darling test

Variables	Pearson χ^2 test		Anderson-Darling test	
	Pearson χ^2 statistic	p-value	A-D statistic	p-value
x11	24.248	0.1471	0.38504	0.3909
x21	51.684	4.183e-05	3.2514,	3.693e-08
x31	56.885	6.454e-06	1.6315	0.000338
x41	19.307	0.3732	0.75192	0.04986
x51	71.969	2.092e-08	4.98	2.437e-12

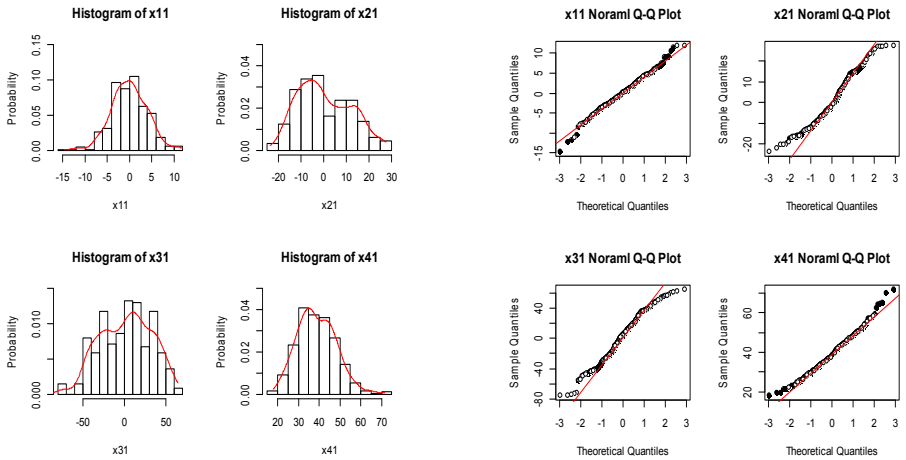

Figure 3. Histogram and normal Q-Q Plot (Normal: solid dots, Abnormal: empty dots)

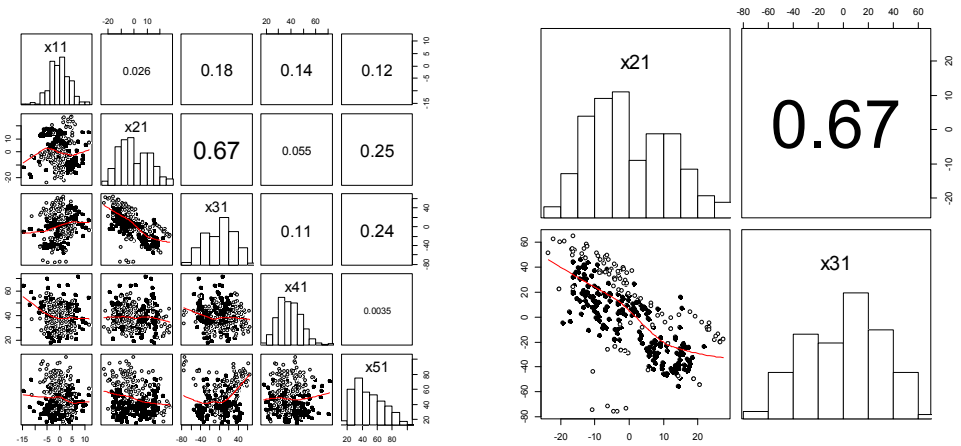

Figure 4. Scatter plots of evaluation data set (Normal: solid dots, Abnormal: empty dots)

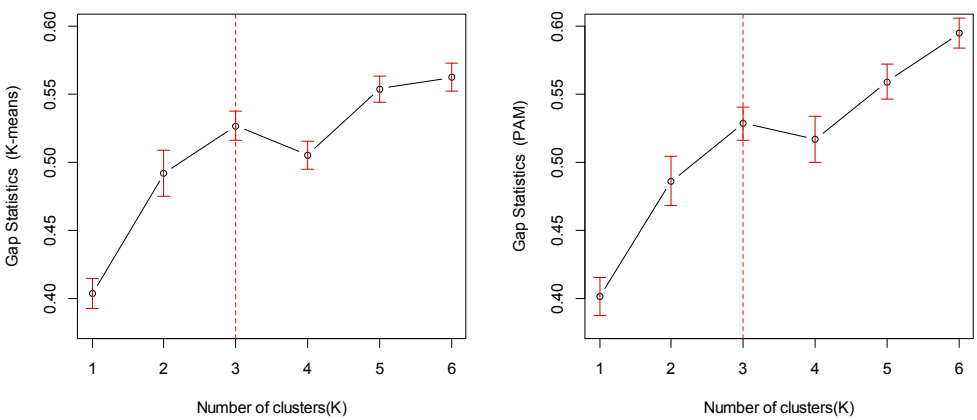

Figure 5. Gap statistics for different number of clusters

Table 5. Determining the best number of clusters(Charrad, et al., 2014)

	All indices of determining the number of clusters				
	KL	CH	Hartigan	CCC	Scott
K	2	2	3	3	3
index	2.643	215.3909	35.0004	24.6209	290.0069
-	Marriot	TrCovW	TraceW	Friedman	Rubin
K	4	3	3	3	3
index	-7.703e+39	1548067737	36245.86	7.6701	-0.3029
-	Cindex	DB	Silhouette	Ratkowsky	Ball
K	3	4	2	4	3
index	0.3212	1.0776	0.3523	0.227	81983.48
-	PtBiserial	McClain	Dunn	SDindex	SDbw
K	4	2	3	4	5
index	0.5989	0.5342	0.1494	0.0646	0.3823

Table 6. Silhouette width and the number of misclassified data(K=3)

Method	Silhouette width		Number of misclassified data
	Average	Abnormal cluster	
K-means_ED	0.29	0.30	13
K-means_ManD	0.29	0.28	13
PAM_ED	0.27	0.27	23
PAM_ManD	0.29	0.33	11

군집분석을 기반으로 설비이상을 탐지하기 위해 제안된 K-means, PAM 수행을 위해서는 모두 군집개수 k를 지정해 주어야 하는데 k를 구하기 위하여 군집수 결정 지표 20가지(참고 : Charrad et al. 2014)에 대한 검토한 결과가 Table 5에 나타나있다. k=2로 판단한 지표가 4개, k=3은 10개, k=4는 5개, k=5는 1개로 나타나 다수결의 원칙(majority rule)에 의거 최적 군집수는 k=3개로 결정하였다. Figure 5는 군집수 결정을 위한 Gap 통계량을 그래프에 도시하였다. 20개의 지표와 마찬가지로 k=3에서 기울기가 급격히 감소하는 그래프를 통해서도 확인 할 수 있다. 군집분석 방법으로는 K-means와 PAM을 사용한 후 그 성능을 비교하였다. 즉, K-means에 유클리드거리(K-means_ED), PAM에 유클리드거리(PAM_ED), K-means에 맨하탄거리(K-means_ManD), PAM에 맨하탄거리(PAM_ManD)의 4가지 알고리즘에 대해 성능을 비교분석하였다. Table 6 은 군집분석 방법별로 군집수 k의 증가에 따른 실루엣 폭(silhouette width)을 나타내었다. Table 6에 사용된 군집화 알고리즘 별 실루엣 값은 Figure 6에 도시되어 있다. 실루엣 값이 클수록 군집내 응집성이 높고, 군집간 분리성이 높아 군집이 잘 나누어진 것으로 판단 할 수 있다. 최적 군집수 k를 3으로 하여 분석해보면 Table 6에서 보는 바와 같이 군집분석 방법별로 평균 실루엣 폭은 큰 차이가 없으나 이상군집에 해당하는 실루엣 폭 값이 PAM_ManD가 0.33로 가장 우수한 성능을 보였다. K-means_ED가 0.30, K-means_ManD가 0.28, PAM_ED가 0.27로 가장 낮은 값을 보였다. Figure 7에 도시된 바와 같이오분류 된 데이터의 갯수도 PAM_ManD가 11개로 가장 작게 나타났고 K-means_ED와 K-means_ManD가 13개, PAM_ED가 23개로 가장 많았다. K-means 방법은 랜덤하게 정해지는 초기치에 따라 성능의 편차가 심하게

나타나 높은 분류성능을 얻기 위해서는 반복적인 초기치 세팅으로 보다 나은 군집을 찾아가는 것이 필요하다. 또한 K-means는 이상치와 잡음의 영향을 많이 받는 반면 대표객체를 사용하는 PAM방법은 K-means 대비 이상치와 잡음에 둔감한 장점을 가진다. 이에 따른 결과로 대표객체를 이용하는 PAM_ManD의 잡음에 대한 둔감성이 가장 우수한 성능을 보인 원인으로 분석된다. 다만, PAM 알고리즘이 K-means에 비해 계산량이 많은 단점을 가지므로 이를 보완하기 위해 의미 있는 입력변수를 선택는 과정(feature selection)과정을 거치면 보다 효율적인 알고리즘이 구성될 수 있을 것으로 보인다. 이때, 본 논문에서 제안한 군집기반 설비상태모니터링 결과를 통해 반응변수(Y)를 라벨링하고 이를 지도학습(supervised learning)을 통해 중요한 변수 선정에 활용할 수 있을 것이다.

군집분석 알고리즘의 성능을 평가할 때 주로 사용하는 시간의 흐름에 무관한 경우 전체 데이터셋을 거리 기반으로 분류하고 의미를 분석하게 된다. 하지만 본 논문에 사용된 설비데이터는 시계열 데이터로써 군집결과와 함께 특정 군집이 나타나는 시점을 탐지하는 목적이 중요하게 고려되었다. 정상상태에서 나타나지 않던 군집이 나타나는 것은 곧 설비이상을 의미하기 때문이다. 또한 비계층적 군집화의 중요한 파라미터인 군집수 K를 20가지 성능지표를 이용하여 다수결의 원칙으로 선정하였다. 본 논문에서 제시한 군집기반 설비이상 탐지절차를 이용하면 설비의 정상상태와 이상상태를 군집화 할 수 있음을 실제 철강제조공정 데이터를 이용하여 확인하였다. 군집분석 기반 알고리즘을 통해 이상군집의 탐지될 경우 현장 설비점검을 실시하도록 함으로써 예측하지 못한 장시간 비계획 정지시간을 감소시킴으로써 설비가동률을 높일 수 있음을 보였다.

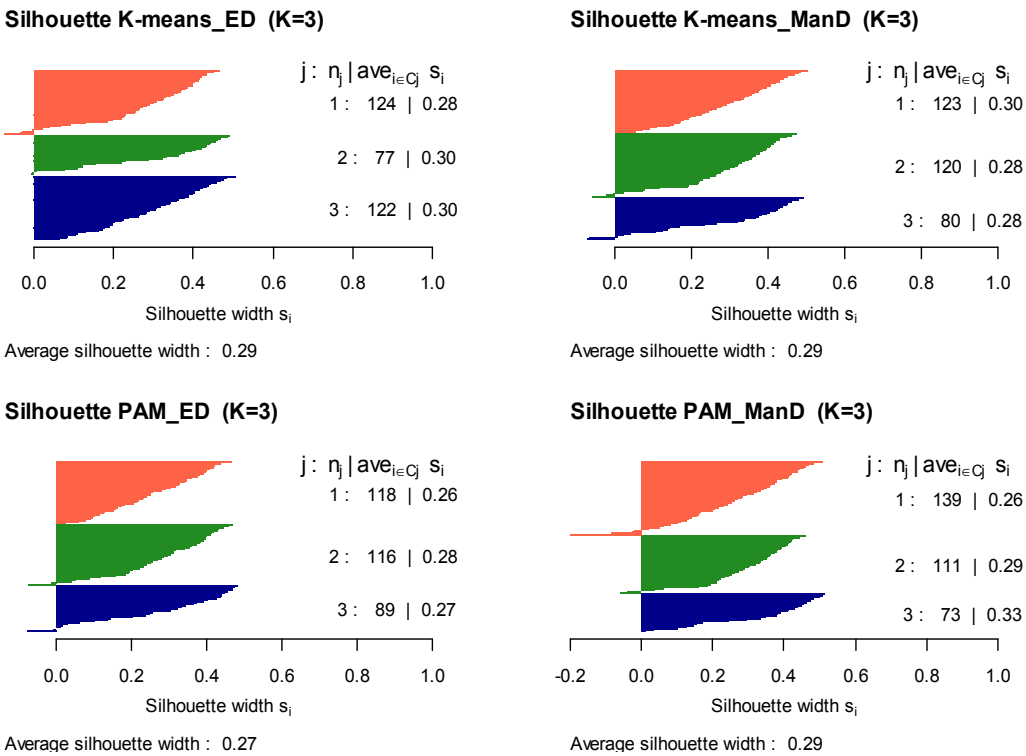

Figure 6. Silhouette plot(K=3)

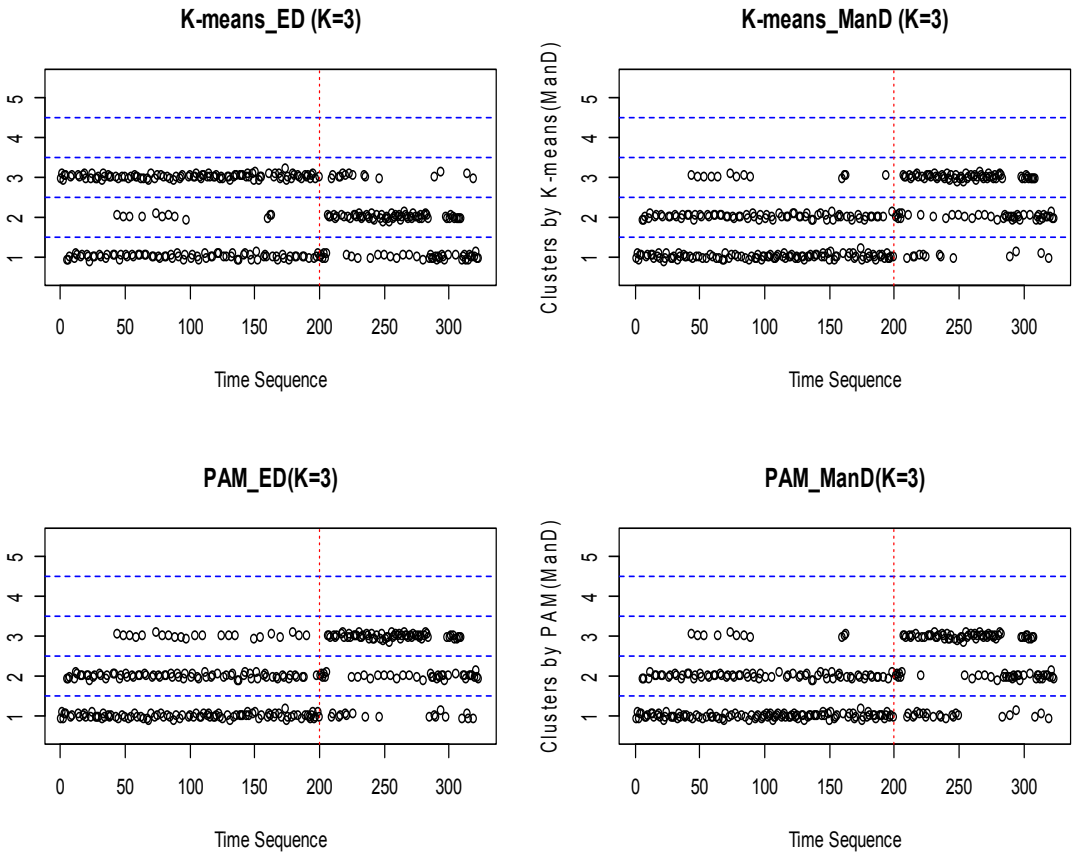

Figure 7. Cluster jitter plot(K=3)

4. 결론 및 추후 연구과제

본 연구에서는 군집기반 설비이상 진단 절차을 제안하였고 현장에서 수집된 데이터를 통해 설비의 비정상 상태를 탐지할 수 있음을 보였다. K-means와 PAM을 이용한 군집분석시 비유사성의 척도로 유클리드거리와 맨하탄거리를 사용하여 성능을 비교하였고, PAM과 맨하탄거리를 결합했을 때 가장 좋은 탐지성능을 보였다. Gap통계량과 20개의 지표를 통해 다수결의 원칙을 통해 최적 군집수를 결정하였다. 또한 다변량공정관리도의 대표적인 방법인 Hotelling T^2 관리도를 사용한 경우 설비이상발생 시점에서만 관리상한(UCL)을 넘는 반응을 보여 지속적인 모니터링과 설비 이상검출에는 한계를 보였다. 실험에 사용된 4종류의 군집기반 알고리즘. 즉, K-means_ED, K-means_ManD, PAM_ED, PAM_ManD 중 PAM_ManD가 군집성능 지표인 최대 실루엣 폭과 최소 오분류 갯수를 나타내어 가장 우수한 성능을 보였다. 이는 대표객체를 사용하는 PAM의 이상치에 둔감한 특성이 반영된 결과임을 확인 할 수 있었다. 향후 연구로는 본 연구에서 사용한 모델을 확장시켜 변수간의 상관관계를 고려하는 거리인 마할라노비스거리(mahalanobis distance)를 이용하는 방법과, 변수선택(feature selection)을 통한 효율적 진단모델을 설계하고자 한다. 또한 국부이상치(local outlier)를 제거하는 정제과정을 통해 잡음에 둔감한 모델을 설계하고 이를 기존의 방법과 비교 평가하는 연구와 필요하고 진행중이다.

REFERENCES

Andrea Cerioli. 2005. "K-means Cluster Analysis and Mahalanobis Metrics: a problematic match or an overlooked opportunity." Statistica Applicata 17:61-73.

Asha Bharambe, Rahul Ravindra, Riya Suchdev, and Yash Tanna. 2014. "A Robust Anomaly Detection System." IEEE International Conference on Advances in Engineering & Technology Research 1-7.

Ghislain Verdier, and Ariane Ferreira. 2011. "Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing." IEEE Transactions Semiconductor Manufacturing 24(1):59-68.

J. E. Kwak, and C. W. Kim. 2013. "Adaptive Clustering Based k-Nearest Neighbor Algorithm for Process Fault Detection." Proceedings of KORMS/KIIE Spring Joint Conference 1169-1175.

J. M. Peña, J. A. Lozano, and P. Larrañaga. 1999. "An Empirical Comparison of Four Initialization Methods for the K-means Algorithm.", Pattern Recognition Letters 20(10):1-17.

Ji Hoon Kang, and Seoung Bum Kim. 2013. "A Clustering Algorithm-Based Control Chart for Inhomogeneously Distributed TFT-LCD Process." International Journal of Production Research 51(18):5644-5657.

Malika Charrad, Nadia Ghazzali, Veronique Boiteau, and Azam Niknafs. 2014. "NbClust: An R Package for Determining the Relevant Number of Clusters in a Dataset." Journal of Statistical Software 61(6):1-36.

R. Gnanadesikan, J. W. Harvey, and J. R. Kettenring. 1993. "Mahalanobis Metrics for Cluster Analysis." The Indian Journal of Statistics Series A(1961-2002) 55(3):494-505.

S. Bersimis, S. Psarakis, and J. Panaretos. 2007. "Multivariate Statistical Process Control Charts: an Overview." Quality and Reliability Engineering International 23(5):517-543.

Sachin Kumar, W. S. Chow, and Michael Pecht. 2010. "Approach to Fault Identification for Electronic Products Using Mahalanobis Distance." IEEE Transactions on Instrumentation and Measurement 59(8):2055-2064.

T. H. Lee, and C. W. Kim. 2013. "Statistical Comparison of Data Mining Models for Fault Diagnosis in an Etching process." Proceedings of KORMS/KIIE Spring Joint Conference 1887-1895.