https://doi.org/10.5392/JKCA.2017.17.07.182

DBSCAN 기반의 제조 공정 데이터 불량 위치의 검출

Detection of the Defected Regions in Manufacturing Process Data using DBSCAN

최은석*, 김정훈*, 아지즈 나스리디노프*, 이상현**, 강정태**, 류관희*
충북대학교 컴퓨터과학*, (주)유라**

Eun-Suk Choi(chl1439@naver.com)*, Jeong-Hun Kim(etyanue@chungbuk.ac.kr)*,
Aziz Nasridinov(aziz@chungbuk.ac.kr)*, Sang-Hyun Lee(lsh0431@yura.co.kr)**,
Jeong-Tae Kang(kangjt@yura.co.kr)**, Kwan-Hee Yoo(khyoo@chungbuk.ac.kr)*

요약

제조 산업은 국가 경제 성장의 원동력으로 그 중요성이 부각되고 있다. 이에 따라 제조 공정상에서 생성되는 제조 데이터 분석의 중요성 또한 조명 받고 있다. 본 논문에서는 PCB(Printed Circuit Board) 제조 공정에서 발생한 로그 데이터를 분석하여 PCB 상에서 빈번하게 발생하는 고장 영역에 대해서 작업자가 고장 영역을 직접 눈으로 볼 수 있도록 시각화하는 방법을 제안한다. 우선 고장 영역을 파악하기 위해서 PCB 공정 데이터 집합에 K-means, DB-SCAN 클러스터링 알고리즘을 적용하여 군집화 하였고, 두 알고리즘 중 더 정확한 고장 영역을 도출하는지 비교하였다. 또한 MVC(Model-View-Controller) 구조 시스템을 개발하여 실제 PCB 이미지 상에 클러스터링 결과를 출력하는 것으로 실제 고장영역을 눈으로 확인할 수 있도록 시각화하였다.

■ 중심어 : | 제조 데이터 | PCB | 고장 영역 |

Abstract

Recently, there is an increasing interest in analysis of big data that is coming from manufacturing industry. In this paper, we use PCB (Printed Circuit Board) manufacturing data to provide manufacturers with information on areas with high PCB defect rates, and to visualize them to facilitate production and quality control. We use the K-means and DBSCAN clustering algorithms to derive the high fraction of PCB defects, and compare which of the two algorithms provides more accurate results. Finally, we develop a system of MVC structure to visualize the information about bad clusters obtained through clustering, and visualize the defected areas on actual PCB images.

■ keyword : | Manufacturing Data | PCB | Defected Region |

I. 서 론

1. 연구의 배경 및 필요성

ICT(Information and Communications Technologies)는 창조경제의 기반으로서 최근에는 빅 데이터[1], 모바일, 웨어러블, IoT가 화두가 되고 있다. 또한 ICT를 활용하고 있는 융합, 복합 산업은 최근 들어 더욱 더 가속화되어지고 있으며, 광범위한 영역으로 우리

* 본 연구는 2016년도 산업통상자원부 지원 사업(사업번호:1005-1028)과 미래창조과학부 및 정보통신기술진흥센터의 대학 ICT연구센터육성 지원사업의 연구결과로 수행되었음"(IITP-2017-2013-0-00881)의 연구결과로 수행되었음
접수일자 : 2017년 04월 14일
수정일자 : 2017년 05월 25일
심사완료일 : 2017년 05월 25일
교신저자 : 류관희, e-mail : khyoo@chungbuk.ac.kr

생활 속에 스며들고 있는 추세이다. 이런 사회적 배경에서 최근 주목 받고 있는 정책은 제조업에 있어 가장 강력한 경쟁력을 가지고 있는 국가 중의 하나인 독일이 시작한 Industry 4.0이라고 할 수 있는데, 독일이 2011년에 발표한 산업 부흥 정책인 '첨단기술전략 2020 실행계획(High-Tech Strategy Action Plan 2020)'에 2012년 추가 공표되어진 정책이다.

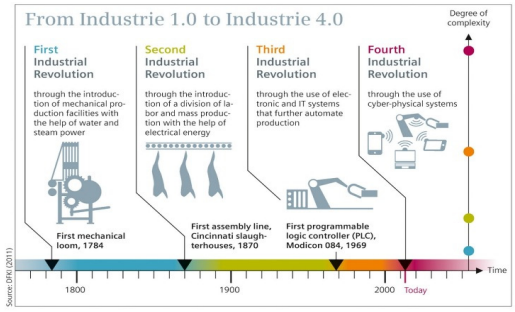


그림 1. 연도별 산업혁명의 포인트[3]

[그림 1]은 18세기 후반의 Industry 1.0부터 최근 Industry 4.0까지의 주요 포인트를 언급하고 있으며 Industry 3.0에서의 IT 기술의 빠른 확산과 컴퓨터 제어를 통한 정보 및 생산 자동화 시스템을 통해 세계적인 부흥을 이뤄냈지만 또 다른 도약을 위하여 Industry 4.0 시대를 선언했다. Industry 4.0은 제조업에 사이버 물리 시스템 (CPS: Cyber Physical Systems)[2], 사물 인터넷 (IoT: Internet of Things)[3], 클라우드 컴퓨팅 (Cloud Computing), 빅 데이터[4] 개념을 내포하고 있으며, 최근 제조현장에서는 스마트폰과 태블릿 PC를 이용한 포터블 기기 간 네트워크 발달을 이용하여 개별 기기를 자율적으로 제어하며 공정의 현 상태나 사물 인터넷을 통해 설비공정에서 발생되어지고 데이터베이스에 저장되고 있는 공정 필수 설비 데이터들을 활용하여 제조 공정의 현황을 실시간으로 모니터링하며 관리 할 수 있는 환경을 사이버 물리 시스템을 통해 구축해 나가고 있으며, 쌓여가고 있는 엄청난 양의 공정 관련 설비 데이터들을 빅 데이터 측면으로 접근해 다양한 방법의 데이터마이닝 기법을 통해 공정 설비 데이터를 분석하여 최적의 공정 설비 설정 값 등을 예측해 주는 분석 모듈[5]을 생성하고, 이를 통해 최소비용 대비 최대수익을 얻기 위한 분석모듈 및 불량이 발생한 제품의 불량 원인을 파악해 다시 같은 불량이 나올 확률을 줄여나가는 분석모듈 적용에 관한 연구가 활발히 진행 중이다. 하지만 이러한 제조 산업의 흐름에서 한국의 제조업도 Industry 4.0에 맞추어 흘러가야 하는데, 현실은 그렇지 못하다. 일부 대기업을 제외한 대부분의 중, 소 제조 기업들의 설비들은 대부분 낙후되어있고 심지어 설비관련 데이터를 데이터베이스화 시키지 않은 제조 기업들도 존재한다. 따라서 제조업의 설비 데이터를 통한 공정 분석연구를 보다 활성화시키기 위해서는 현장상황 개선이 필요하다.

2. 연구의 목표

본 연구는 데이터마이닝 클러스터링 기법들을 사용하여 현 제조업 설비 데이터에 적합한 클러스터링 방법을 찾기 위해 성능을 비교분석한다. DBSCAN (Density-based spatial clustering of applications with noise), K-means 알고리즘을 사용해 제조 프로세스 상의 PCB (Printed Circuit Board)[6] 에 컴포넌트들을 장착하는 공정과 컴포넌트 장착이 올바른 위치에 있는지, 외관 상황을 파악하는 AOI(Automatic Optical inspection) 공정의 설비데이터들을 확보하고 PCB에 장착한 각 컴포넌트들의 좌표 데이터를 활용해 불량이 일어난 데이터들을 군집화 및 가시화해 불량 컴포넌트의 위치를 작업자들이 쉽게 파악 할 수 있도록 하는 것이 목표이며 연구의 진행 순서는 다음과 같다.

- 사용하게 될 알고리즘에 알맞은 데이터를 갖고 있는 공정을 선택해 PCB 식별자(ID)와 제품의 불량 유무를 고려해 각 컴포넌트 별로 데이터 집합을 확보한다. 컴포넌트의 좌표 데이터가 없는 설비 데이터 테이블에서는 좌표 데이터를 가지고 있는 설비 데이터 테이블에서 같은 컴포넌트 이름을 가진 설비 데이터를 통해 해당 컴포넌트의 좌표 데이터를 확보한다.
- 선택된 공정들의 완성된 데이터 집합을 사용해 DBSCAN, K-means 클러스터링 알고리즘을 사용해 서로 비교분석한다.

- 가장 적합하다고 판단되는 알고리즘을 선택해 PCB 이미지에 매핑 시켜 실제로 어느 부분에서 불량 군집이 형성되는지 확인 할 수 있도록 MVC 기반의 시스템을 구축해 가시화한다.

II. 관련 연구

1. DBSCAN(Density-based spatial clustering of application with noise)

DBSCAN(Density-based spatial clustering of applications with noise)[7]은 Martin Ester 외 3명이 제안한 클러스터링 방법으로 기존의 데이터간의 거리를 파악해 클러스터를 형성하는 알고리즘과는 달리 같은 군집 내의 각 데이터들은 서로 밀도가 높게 형성되었을 것이라는 가정 하에 각 데이터 포인트 주변의 밀도를 이용해 클러스터를 형성시키는 알고리즘이다. 이 알고리즘에서는 두 개의 파라미터 정의가 있어야한다. 첫 번째로 주변 공간에 대한 정의가 필요하다. 두 번째로는 주변 공간 내부에 최소 몇 개의 데이터 포인트가 존재해야 클러스터로 인정 할 것인지에 대한 정의가 필요하다. 각 데이터 포인트들로부터의 거리 반경을 epsilon(ϵ)라고 칭하며, 클러스터로 인정하기 위해 필요한 최소의 데이터 포인트 개수는 minPts(τ)로 칭한다. 이와 같은 파라미터를 사용해 아래의 개념을 정의 할 수 있다.

- 이웃 포인트(neighborhood of a point): 한 데이터 포인트로부터 반경 내에 존재하는 다른 데이터 포인트를 이웃 포인트라 정의함.
- 코어 포인트(core point): 최소의 데이터 포인트인 minPts개 이상의 이웃 포인트를 갖는 데이터 포인트를 코어 포인트라 정의함.
- 직접 접근 가능한 포인트(directly density reachable): 어떤 코어 포인트 p의 이웃 포인트 q에 대하여 코어 포인트 p는 이웃 포인트 q에 직접 접근 가능하고 이를 $(p \rightarrow q)$라 정의함.
- 접근 가능한 포인트(density-reachable): 데이터 포인트 p, q 에 대하여 직접 접근 가능한 포인트 배열 $p = \{p_1, p_2, p_3, ..., q\}$이 존재한다면 q는 p로부터 접근 가능한 포인트로 정의함.
- 연결된 포인트(density-connected): 데이터 포인트 p, q에 대하여 접근 가능한(density-reachable) 데이터 포인트 o가 존재한다면 데이터 포인트 o는 p, q와 연결된 포인트라고 정의함.
- 클러스터(cluster): 하나의 코어 포인트 p에 대하여 접근 가능한 포인트(density-reachable)들의 집합을 클러스터로 정의하며, 하나의 클러스터 내의 모든 데이터 포인트들은 서로 연결된 포인트(density-connected)이다.
- 노이즈(noise): 어떠한 클러스터에도 속하지 않는 데이터 포인트를 노이즈로 정의한다.

DBSCAN은 다음과 같은 단계로 수행된다.

- 알고리즘에 사용될 데이터베이스에서 코어 포인트의 조건을 만족시키는 점을 임의로 선택하여 seed로 선별한다.
- 선별된 seed를 기준으로 접근 가능한 포인트(density-reachable)들을 모두 찾아 클러스터로 인식시킨다.
- 위의 단계를 모든 데이터 포인트를 대상으로 반복한다.

정의 되어진 각 클러스터들은 연결된 포인트(density-connected) 조건을 만족하며, 거리 반경 epsilon 이내에 최소 데이터 포인트 개수 minPts 이상의 개수를 갖는다.

2. K-means++

K-means++[8]는 Arthur. D외 1명이 2007년 제안한 클러스터링 방법으로, 본 연구에서 활용하는 분할 클러스터링 알고리즘 중 가장 빈번히 사용되어지는 방법이다. 비지도 학습의 대표적인 모델로서 모집단 또는 범주에 대한 사전정보가 없는 경우 주어진 데이터 포인트들 사이의 거리 측정 및 유사성을 이용해 분석을 하며, 전체 데이터를 K개의 집단으로 그룹화 시켜 각 집단의 성격을 파악해 나감으로써 전체 데이터 집합의 구조에

대한 이해를 돕기 위해 사용한다. K-means가 수행되는 과정은 다음 [그림 2]와 같다.

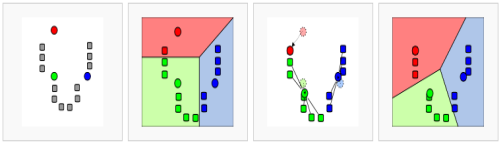


그림 2. K-means 진행 순서도

K-means는 분석과정 초기에 클러스터의 수를 결정하고 동일한 수만큼의 중심점(Centroid)을 임의로 설정한다. 생성한 중심점을 기준으로 각 데이터 포인트를 가장 가까운 클러스터에 할당시키는데 이때 중심점과 각 데이터 포인트 간의 거리 계산은 Euclidean distance로 계산하게 된다. 이후 각 클러스터의 중심점을 다시 계산하고 앞의 과정을 반복한다. 반복하는 과정에서 새로 계산된 중심점이 이전 단계의 중심점과 같다면 반복을 중단하고, 다르다면 계속해서 반복 수행한다.

III. DBSCAN을 사용한 불량 위치 검출 시스템

제 3절에서는 제조기업의 공장 설비데이터를 활용해 PCB(Printed Circuit Board)에 장착된 각 컴포넌트들의 불량 집중 형성 구역을 판별하기 위한 DBSCAN을 사용한 불량 위치 검출 시스템을 제안한다. 현재 한국의 제조 산업에서는 불량률 관리나 생산율 최적화 관리에 많은 관심을 쏟고 있지만 제품을 작동시키는 핵심적인 역할을 하는 PCB의 컴포넌트들의 주요 불량 발생위치와 빈도에 대한 시각 고도화는 이루어지지 않고 있다. 본 논문에서는 제조 데이터에 대한 사전지식과 데이터 분석에 대한 전문지식을 가지지 못한 사용자에게 시각 고도화된 분석 결과를 제공함으로써 PCB에 장착된 컴포넌트들의 주요 불량발생 위치와 빈도를 쉽게 파악해 우발적인 상황에 즉각적인 대처를 할 수 있도록 한다. 따라서 현재 제조업에서 PCB 컴포넌트 장착 공정 및 컴포넌트 장착이 올바른 위치에 있는지, 외관 상황을 파악하는 AOI (Automatic Optical inspection) 공정들의 설비 데이터에 포함된 PCB에 장착된 컴포넌트들의 좌표 데이터 집합을 대상으로 데이터 마이닝 기법 중 하나인 클러스터링을 수행하고 결과를 가시화[10] 하는 시스템을 설계하였다. 이를 위해서 시스템에 가장 적합한 클러스터링 방법을 선택하기 위한 연구 또한 병행하였다. 3.1절에서는 본 논문에서 제안하는 시스템 흐름도와 동작원리를 다룬다. 3.2절에서는 클러스터링 방법을 선택하기 위한 실험에서 사용할 데이터 집합을 정의하고 설명한다. 마지막 3.3절에서는 데이터 집합에 클러스터링 알고리즘을 적용한 결과를 비교분석한다.

1. 시스템 흐름도

실제 제조기업의 공장 설비데이터를 활용해 PCB에 장착된 각 컴포넌트들의 불량 집중 형성 구역을 판별하기 위해 제안하는 방법인 DBSCAN을 사용한 불량 위치 검출 시스템은 [그림 3]과 같이 구성된다.

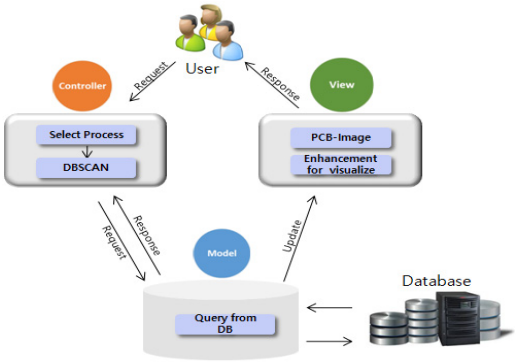


그림 3. 불량 위치 검출 시스템 흐름도

시스템은 우선 사용자로부터 분석 대상 공정을 입력받게 된다. 이후 Controller에서 분석 대상 공정의 설비 데이터를 이용하여 DBSCAN 클러스터링을 수행한다. 본 논문의 불량 위치 검출 시스템에서 DBSCAN을 선택한 이유는 3.3절에서 자세히 다룬다. 분석 대상 공정의 설비 데이터는 사용자의 요청에 대한 질의를 Model에서 Database에 직접 접근하여 처리하고 반환된 데이터 집합이다. 반환된 데이터 집합으로 DBSCAN을 적용한 결과를 View에서 실제 PCB 이미지에 불량 위치를 시각화한다. 이로써 본 논문에서 제시한 시스템은

오직 사용자로부터 특정 제조공정을 입력받는 것만으로 해당 공정의 PCB에서 빈발적으로 발생한 불량 위치를 시각화하여 보여준다는 것에 의미를 가진다고 할 수 있다.

2. 실험데이터 집합 정의

클러스터링 알고리즘에 사용될 데이터 집합은 실제 제조공장에서 수집되어진 공정 설비 데이터로 대상 제조업체에서는 설비 데이터, LEGACY(POP, MES, ERP), 센서(공정 및 창고 내/외부 온/습도), 기상청 정보 총 5가지 부분에 대해서 기본적인 데이터베이스를 구축하고 있으며, 본 연구에서 실질적으로 사용된 데이터 집합은 설비 데이터 중에서 PCB와 연관성이 짙은 2개의 공정을 대상으로 분석을 진행했다. 사용된 2개의 공정은 각 A공정-detail 테이블, A공정-header 테이블, B공정-detail 테이블, B공정-header 테이블을 클러스터링 알고리즘에 적용시켰다. 다른 항목의 데이터베이스 테이블은 분석과정에서 제외되었다. 데이터 수집은 [표 1]과 같이 일정 수집주기를 설정하였다.

표 1. 제조공정 데이터 수집주기

구분	수집주기
설비데이터	1시간
LEGACY(POP,MES,ERP)	1시간
PaperLess	일 2회
센서(공정 및 창고 내/외부 온/습도)	실시간
기상청정보	1시간

[표 2]는 실제 제조업에서 사용되고 수집되어진 PCB 외관 검사 공정의 detail 데이터 속성으로 15개의 속성이 있지만 본 연구에서 제안하는 방법에 사용할 실제 데이터 속성은 데이터 전처리 과정을 거쳐 5개로 추려졌다. 1열은 PCB 제품의 식별을 위한 개별코드를 나타내며, 2열은 작업이 기록된 날짜, 3열은 각 제품 코드별 컴포넌트의 이름을 의미하며, 4열, 5열은 해당 컴포넌트의 PCB 보드 상의 위치 데이터를 의미한다. 하지만 해당 컴포넌트의 불량 유무를 판단할 수 있는 검사 결과 데이터는 해당 테이블에 존재하지 않기 때문에 동일한 공정의 header 테이블까지 한 묶음으로 사용한다.

표 2. PCB연관 1번 공정 detail 데이터 테이블

distinguish	date	component	x	y
Barcode1	2015-12-01	Q14	2633	10991
Barcode2	2015-12-01	TVS5	4259	13978
Barcode3	2015-12-01	U19	2063	9953
Barcode4	2015-12-01	R44	8546	3931
Barcode5	2015-12-01	C89	3109	4689
Barcode6	2015-12-01	C8	5718	12049
Barcode7	2015-12-01	C25	4543	12942
Barcode8	2015-12-01	C39	3688	13862
Barcode9	2015-12-01	C17	5029	15261
Barcode10	2015-12-01	C62	4642	4608

표 3. PCB연관 1번 공정 header 데이터 테이블

distinguish	result	date
Barcode_1	OK	2016-01-08
Barcode_2	OK	2016-01-28
Barcode_3	OK	2015-12-01
Barcode_4	NG	2015-12-24
Barcode_5	NG	2015-11-25
Barcode_6	NG	2015-11-25
Barcode_7	NG	2015-11-26
Barcode_8	NG	2015-11-26
Barcode_9	OK	2015-12-03
Barcode_10	OK	2015-12-16

[표 3]은 PCB 외관 검사 공정의 header 테이블의 속성 9개 중 데이터 전처리를 통해 [표 2]의 각 제품코드별 C컴포넌트의 불량 유/무를 파악하기 위한 최소한의 속성 3가지만 추출하였다. 마찬가지로 1열은 PCB 제품의 식별을 위한 개별코드를 나타내며, 2열은 불량 유무를 나타내는 결과 필드이다. 마지막 3열은 해당 작업이 수행된 날짜를 의미한다.

[표 2]는 제조업의 PCB 검사 연관 공정의 두 번째 항목인 detail 데이터 속성들로서 31개의 데이터 속성이 존재하지만, 본 연구에서 제안 하는 방법에 사용하는 최적화된 데이터를 얻기 위해서 데이터 전처리 과정을 거쳐 3가지의 속성만을 추출하였다. 전 처리된 공정 데이터 속성 중에 컴포넌트의 좌표를 나타내는 속성은 존재하지 않는다. 하지만 제조 공장의 경우 공정은 다르더라도 모든 공정에서 생산하는 PCB의 종류가 같을 경우 같은 좌표 값을 가지기 때문에 각 공정에서 생산되는 각 컴포넌트들은 서로 동일한 x, y 좌표를 가진다고 할 수 있다. 따라서 첫 번째 분석 공정인 PCB 외관 검사 공정의 [표 2] 데이터 속성 중 각 컴포넌트들의 위치

를 나타내는 x, y 데이터를 이용해 해당 공정의 컴포넌트들을 파악할 수 있다.

표 4. PCB연관 2번 공정 detail 데이터 테이블

distinguish	date	component
Barcode_1	2015-10-04	C1
Barcode_2	2015-10-04	C2
Barcode_3	2015-10-04	C3
Barcode_4	2015-10-04	C5
Barcode_5	2015-10-04	C6
Barcode_6	2015-10-04	C12
Barcode_7	2015-10-04	C13
Barcode_8	2015-10-04	C14
Barcode_9	2015-10-04	C15
Barcode_10	2015-10-04	C16

[표 5]는 PCB 연관 검사 두 번째 공정의 header 테이블의 속성 9개 중 [표 4]의 각 제품코드별 컴포넌트의 불량 유/무를 파악하기 위한 데이터만을 얻기 위해 데이터 전처리 과정을 거쳐 속성을 3가지만 추출하였다.

표 5. PCB연관 2번 공정 header 데이터 테이블

distinguish	result	date
Barcode_1	NG	2015-10-04
Barcode_2	NG	2015-10-04
Barcode_3	OK	2015-10-05
Barcode_4	OK	2015-10-04
Barcode_5	OK	2015-10-04
Barcode_6	OK	2015-10-04
Barcode_7	OK	2015-10-04
Barcode_8	OK	2015-10-04
Barcode_9	OK	2015-10-04
Barcode_10	OK	2015-10-04

본 연구에서는 이러한 데이터를 기반으로 제조공정의 PCB Component들을 장착해주는 공정 및 Component가 올바른 위치에 장착되었는지, 외관 상황을 파악하는 공정을 위해 비지도 학습 클러스터링 알고리즘인 K-means, DBSCAN을 적용하여 어느 알고리즘이 실무자들에게 더욱 효율적이고 정확하게 PCB Component들의 불량 위치를 나타내는지 실험한다.

3. 클러스터링 알고리즘 적용 및 실험

본 실험에서는 주어진 제조공정의 PCB Component 위치 및 불량 여부 데이터 집합에 K-means, DBSCAN 클러스터링 알고리즘 적용하여 어떤 알고리즘이 더 정확하고 효율적인 클러스터링 결과를 도출하는지 비교하였다.

두 클러스터링 알고리즘을 수행하는데 있어 데이터 집합에서 사용하는 속성은 x, y, component 세 가지 속성이다. 거리 계산 방법은 유클리디언 거리(Euclidean distance)[9]과 맨하탄 거리(Manhattan distance)[11] 중 데이터 포인트 사이의 직선거리를 측정하는 유클리디언 거리 측정 방법을 사용했으며, 유클리디언 거리를 구하는 방법은 식 1과 같다.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

여기서 p, q는 임의의 두 점이며, p_i와 q_i는 각각 두 점을 구성하는 좌표 값이다. $d(p,q)$는 수식에 대한 결과로 두 점 p, q 사이의 거리에 해당한다.

다음 [그림 4]는 실험에 사용될 제조업 PCB 관련 1번 공정의 데이터 집합을 산점도에 표현한 결과이며, [그림 5]는 실험에 사용될 제조업 PCB 관련 2번 공정의 데이터 집합을 산점도에 표현한 결과이다.

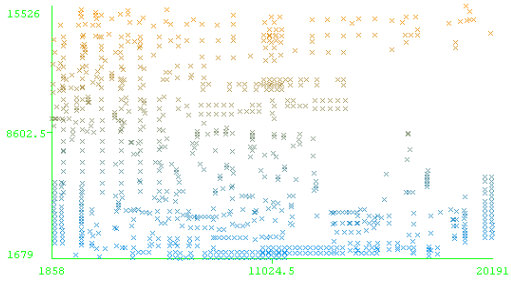


그림 4. PCB 1번 공정 데이터 집합 산점도

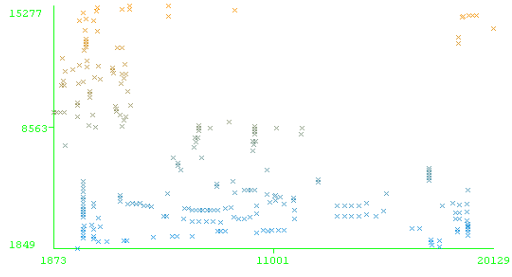


그림 5. PCB 2번 공정 데이터 집합 산점도

첫 번째 실험은 PCB공정 데이터 집합에 K-means 알고리즘을 적용하여 불량 위치를 검출한다. K-means 알고리즘을 수행하기 위해서는 형성할 클러스터의 개수를 사용자가 지정해야한다. 본 연구에서는 K 값을 3으로 고정하고, 데이터의 집합을 10000개에서 20000개로 변화를 주면서 각각의 데이터 집합에 대해 실험을 진행하였다.

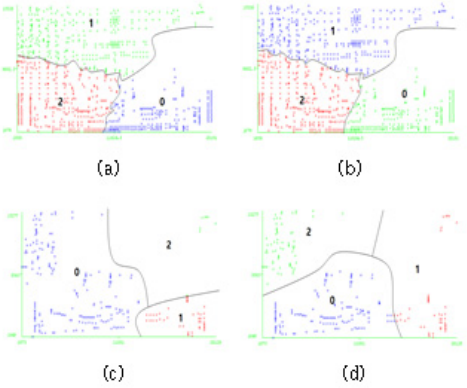


그림 6. K-means 클러스터링 결과

위의 [그림 6]은 K-means 알고리즘의 파라미터 K를 3으로 설정하여 클러스터링을 수행한 결과이다. 그 중 (a)와(c)는 데이터 집합의 레코드가 10000개인 경우이고, (b)와 (d)는 데이터 집합의 레코드가 20000개인 경우이다. (a), (b)는 제조업의 PCB관련 1번 공정에 해당하는 클러스터링 결과이다. 그림에서 확인할 수 있듯이 데이터양의 증가에 따른 클러스터 형성에는 큰 변화가 없었다. 하지만 PCB 2번 공정에 해당하는 (c), (d)의 경우 생성된 클러스터의 모양에 변화가 생기게 되는데 이는 K-means 알고리즘이 임의의 초기 중심점 위치에 큰 영향을 받고, 항상 일관된 클러스터를 형성하지 않는다는 것을 알 수 있다. 또한 (a), (b), (c), (d) 모두 전체 데이터 객체를 대상으로 군집을 형성한 것으로 보아 K-means 알고리즘의 경우 노이즈 데이터를 구분할 수 없기 때문에 PCB상의 불량 위치를 검출하는 데에는 부적합하다는 것을 알 수 있다.

두 번째 실험은 PCB 공정 데이터 집합에 DBSCAN 알고리즘을 적용하여 불량 위치를 검출한다. DBSCAN의 경우 K-means 알고리즘과 달리 형성할 클러스터의 수를 미리 지정할 필요가 없다. 대신 코어 포인트로부터의 반경 값인 epsilon과 반경 내의 최소 데이터 포인트 개수 값인 minPts를 설정해야한다. 본 연구에서는 총 10000개의 공정 데이터를 가지는 PCB 공정 데이터 집합에 epsilon 값을 0.09로 고정하고, minPts 값을 각각 200, 280으로 변경하며 실험을 진행하였다. 다음 [그림 7], [그림 8]은 DBSCAN 알고리즘을 수행한 결과이다.

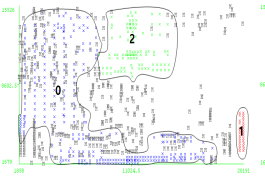


그림 7. ϵ=0.09, τ=200

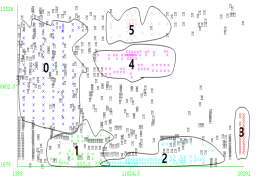


그림 8. ϵ=0.09, τ=280

[그림 7]의 경우 임의의 데이터 포인트의 반경 epsilon 안에 접근 가능한 포인트들이 최소 200개 이상 포함 되어있는 클러스터가 3개 형성되었다. 여기서 클러스터에 포함되지 않은 데이터 포인트는 노이즈로 판별한다. [그림 8]은 minPts 값이 280일 때의 결과로 6개의 클러스터가 형성되었다. [그림 7]의 클러스터 0을 보면 임의의 데이터 포인트에서 접근 가능한 데이터 포인트 개수가 200개 이상인 것들을 찾아 클러스터의 크기를 계속 증가시키기 때문에 비교적 거대한 형태의 클러스터를 형성시킨다. 또한 형성된 클러스터의 의미는 클러스터가 형성되어져 있는 위치에 불량 컴포넌트 데이터 포인트가 밀집 되어 있다는 뜻으로 이는 PCB 관련 공정의 주요 불량 위치 판별에 K-means 알고리즘 보다 적합하다는 의미로 볼 수 있다.

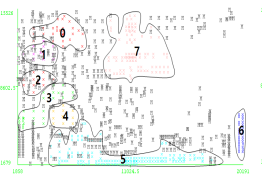


그림 9. ϵ=0.08, τ=200

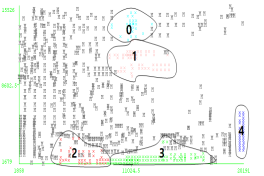


그림 10. ϵ=0.08, τ=280

[그림 9], [그림 10]은 위와 동일한 실험 환경에서 임의의 데이터 포인트의 반경 epsilon 값을 줄인 실험의 결과이다. epsilon 값이 줄어들게 되면 반경 안에서 접근 가능한 포인트들의 개수도 줄어들게 된다. 따라서 [그림 10]과 같이 형성된 클러스터의 수가 줄어든 모습을 볼 수 있다. [그림 9]의 경우 [그림 7]에 비하여 epsilon 값이 작아졌기 때문에 더 작은 범위를 기준으로 접근 가능한 포인트를 계산하므로 하나였던 클러스터가 여러 형태로 나누어 진 모습을 실험을 통해 확인 할 수 있다.

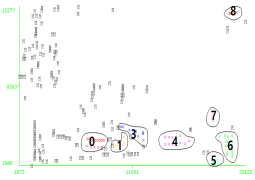


그림 11. ϵ=0.7, τ=250

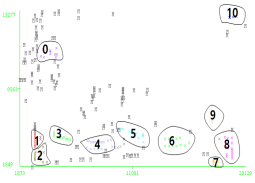


그림 12. ϵ=0.7, τ=200

[그림 11], [그림 12]는 제조업의 PCB 관련 2번 공정 데이터에 대한 DBSCAN 알고리즘 수행 결과이다. 데이터 집합의 밀집도와 분포도가 1번 공정 데이터 집합과 상이하기 때문에 epsilon 값과 minPts 값을 1번 공정의 실험과 동일하게 설정할 경우 클러스터링이 원활하게 수행 되지 않는다. 따라서 [그림 11], [그림 12]에서는 epsilon 값을 0.7, minPts 값을 각각 250, 200으로 설정하고 실험을 진행하였다. 1번 공정의 데이터 집합과 분포 형태가 다르기 때문에 클러스터의 형태 또한 다르게 형성되었다. [그림 12]는 동일한 epsilon 값에 minPts 값을 감소시킨 실험 결과로서 [그림 11]에 비하여 형성된 클러스터의 수가 많다.

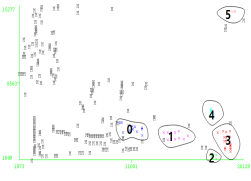


그림 13. ϵ=0.65, τ=250

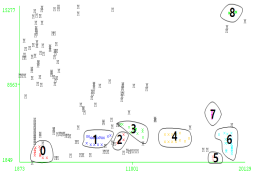


그림 14. ϵ=0.65, τ=200

[그림 13], [그림 14]는 epsilon 값을 0.65로 변경하여 수행한 결과이다. [그림 13]의 경우 epsilon 범위가 줄어들어 반경 안에서의 접근 가능한 포인트들의 개수도 줄어들게 되므로 [그림 11]과 비교 했을 때 클러스터의 수가 하나 줄어들었다. [그림 14]도 마찬가지로 [그림 12]과 비교 했을 때 클러스터의 수가 줄어든 것을 확인 할 수 있다.

이러한 실험 내용을 통해서 알 수 있는 사실은 K-means는 분할성이 좋지만 중심점에 멀리 떨어져 있거나 노이즈 데이터를 판별하는 데는 좋지 않다. 반면 DBSCAN은 K-means 보다 분할성은 떨어지는 편이지만 노이즈 데이터에 강하다는 사실을 실험을 통해 입증되었다. 따라서 현 제조업의 PCB 검사 관련 공정의 불량 컴포넌트 확보 및 불량 밀집 구역 데이터 확보에는 DBSCAN 알고리즘이 더 효과적이라고 할 수 있다.

IV. 실험 결과

1. 불량 밀집 구역 시각 고도화

본 연구에서 구현된 결과는 두 가지 제조공정 데이터 중 PCB 연관 공정에 대한 불량 밀집구역 가시화를 [그림 3]의 제안하는 시스템을 기반으로 구현했으며, GUI 환경에서 사용자가 공정을 선택할 수 있도록 구현 하였다. 결과는 [그림 15], [그림 16]과 같다.

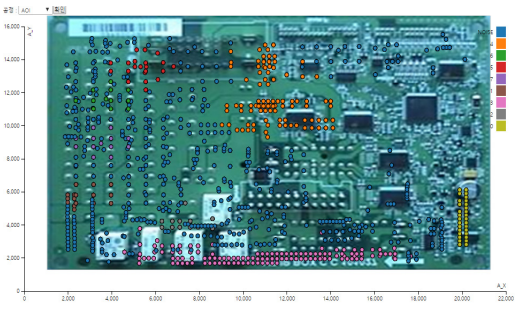


그림 15. PCB 1공정 시각 고도화

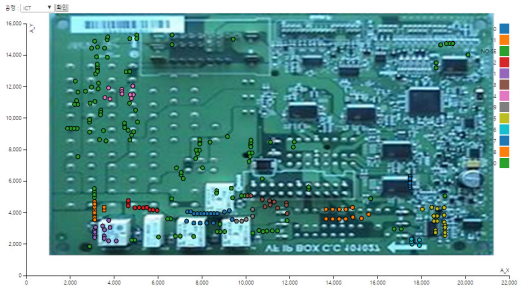


그림 16. PCB 2공정 시각 고도화

[그림 15]는 PCB 1공정 데이터를 분석한 결과를 시각화한 것이다. 실제 PCB 이미지 상에 시각화된 결과를 통해 사용자는 특정 빈도 이상으로 컴포넌트 불량이 발생하는 위치를 특정하여 확인할 수 있다. 이때 빈도는 DBSCAN을 수행하기 위해 입력한 매개변수에 영향을 받는다. 즉, DBSCAN에서 클러스터를 형성하기 위한 밀도 역치 값인 두 개의 매개변수 epsilon과 minPts가 불량 발생 빈도 역치라고 할 수 있다. [그림 15]와 [그림 16]에서 파란색 점을 제외한 나머지 점이 DBSCAN에 의해 형성된 클러스터이며 PCB 컴포넌트 불량이 자주 발생하는 위치이다. 이로써 PCB 좌표 데이터와 통계 데이터에 의존적이었던 이전 PCB 불량관리 방식과 달리 오직 분석을 원하는 PCB 공정을 선택하는 것만으로 해당 공정에서 생산하는 PCB의 불량 발생 위치를 시각적으로 확인하고 나아가 불량이 자주 발생하는 위치와 영역까지 확인 할 수 있다. 다음 4.2절에서는 DBSCAN을 적용하기 위해 가장 적합한 epsilon과 minPts를 결정하기 위한 실험결과를 다룬다.

2. 입력 매개변수 결정

본 연구에서는 제조업의 여러 공정 중 제품 작동에 핵심적인 역할을 하는 PCB의 각 컴포넌트 데이터를 활용해 불량이 자주 일어나며, 불량 밀집도가 높은 위치를 판별하고 가시화하기 위하여 잘 알려진 클러스터링 방법들인 K-means, DBSCAN을 사용하였다. 클러스터링을 위한 최적의 매개변수를 결정하기 위해서 클러스터링 결과 품질척도 중의 하나인 균질도(Homogeneity)와 알고리즘 수행 속도를 사용하였다. 균질도란 두 개의 데이터 포인트 u와 v 사이의 유사도를 두 데이터 포인트의 상관 계수로 정의하며 두 개의 데이터 포인트가 양의 상관계수를 갖는다면 1에 가까운 값을 가지고, 음의 상관계수를 갖는다면 -1에 가까운 값을 가지며, 상관관계가 적을수록 0에 가까운 값을 갖는다. 상관계수를 측정할 거리 계산 알고리즘은 Euclidean distance를 사용했으며, 각 클러스터들의 중심점들과 해당 클러스터 안에 속해있는 포인트들 간의 평균 거리로 계산하게 된다[12]. 동일 클러스터 안에 포함되어 있는 각 데이터 포인트들의 상관계수는 높아야 하며 균질도가 높을수록 클러스터링 결과가 좋다고 판단 할 수 있다. 균질도의 계산식은 다음과 같다.

$$H_{ave} = \frac{1}{N_{point}} \sum_{i} D(p_i, C(p_i)) \qquad (2)$$

N_{point}는 전체 데이터 포인트의 수, D는 거리 계산 알고리즘, p_i는 i번째 데이터 포인트, $C(p_i)$는 p_i가 속해있는 클러스터의 중심점을 의미한다.

표 6. 성능평가 결과 표

알고리즘	파라미터	수행시간	균질도
DBSCAN 1공정	0.09, 280	3.57s	0.276
DBSCAN 1공정	0.09, 200	3.49s	0.240
DBSCAN 1공정	0.08, 280	3.29s	0.261
DBSCAN 1공정	0.08, 200	3.41s	0.285
DBSCAN 2공정	0.7, 250	3.15s	0.180
DBSCAN 2공정	0.7, 200	3.28s	0.181
DBSCAN 2공정	0.65, 250	3.15s	0.163
DBSCAN 2공정	0.65, 200	3.23s	0.176
K-means 1공정	3	0.08s	0.068
K-means 1공정	5	0.13s	0.084
K-means 2공정	3	0.07s	0.032
K-means 2공정	5	0.09s	0.038

[표 6]은 실험 단계에서 수행했던 K-means, DBSCAN 알고리즘 데이터 집합을 활용하여 클러스터링 결과 성능 판별과 실제 데이터에 적용하기 위해 각 알고리즘별, 각 공정별, 각 파라미터 별로 나누어 균질도와 수행시간을 측정했다. 결과적으로 1공정에서의 DBSCAN 알고리즘은 epsilon 값이 0.08이며 minPts 값이 200이었을 때 균질도가 가장 높게 나타나 이를 실제 1공정 결과에 적용시켰다. 2공정에서의 DBSCAN 알고리즘은 epsilon 값이 0.7이며, minPts 값이 200인 경우에 균질

도가 가장 높게 나타났고 이를 2공정 결과에 적용시켰다. K-means 알고리즘의 경우 노이즈 데이터를 식별할 수 없고, 불량 밀집구역을 식별 할 수 없었기 때문에 결과 적용에서 제외하였다.

V. 결론

최근의 가장 큰 이슈인 ICT 융합 그리고 Industry 4.0이 가져오고 있는 변화에 발맞추어 나갈 수 있도록 제조 공정 기반의 데이터 분석 연구를 진행했다. 모든 전자기기에 필수적으로 들어가게 되는 PCB 관련 공정 데이터를 활용한 분석 방안은 아직 국내에 많이 연구된 바가 없다.

본 연구에서는 DBSCAN 기반의 제조 공정 데이터의 불량 위치의 검출에 대한 방법을 제안했다. K-means와 DBSCAN 클러스터링 알고리즘을 통해 제조업 PCB 관련 2공정 데이터를 기반으로 어떤 클러스터링 알고리즘이 PCB의 불량 위치를 더 효과적으로 분할하고, 불량 밀집 구역을 더 잘 표현할 수 있는지 실험하였고, 가장 최적의 클러스터를 형성하기 위한 파라미터 값을 결정하기 위한 성능평가 또한 실험을 통해 수행하였다.

성능평가 기준인 균질도 측정에서 DBSCAN 알고리즘이 K-means 알고리즘보다 분석결과의 상관관계 및 성능이 우수하다고 평가 되었으며, MVC 기반의 시스템 구조에 적용시켜 실 사용자들이 쉽고 간편하게 PCB 관련 공정의 주요 불량 밀집도를 한 눈에 확인 할 수 있도록 시스템을 구축했다.

향후 연구로는 DBSCAN 알고리즘의 파라미터인 데이터 포인트의 반경 epsilon과 클러스터로 인정하기 위한 반경 내의 최소 데이터 포인트 개수 minPts를 공정 데이터의 형태와 사이즈에 따라서 자동으로 최적치를 추천해주는 알고리즘을 연구하도록 하겠다.

참 고 문 헌

[1] 최종우, 이일우, “빅 데이터를 활용한 제조공정 결함 예측에 관한 연구,” 한국데이터정보과학회지, 제26권, 제5호, pp.1141-1154, 2015.

[2] K. KOBARA, "Cyber Physical Security for Industrial Control Systems and IoT," IEICE Trans. Inf. & Syst., Vol.99, No.4, pp.787-795, 2016.

[3] A. Pühringer, “사물인터넷(IoT)과 인더스트리 4.0(Industry 4.0)이 기존의 산업용 통신에 미치는 영향,” ICN, pp.22-27, 2015.

[4] J. Lee, E. Lapira, B. Behrad, and K. Hung-an, "Recent advances and trends in predictive manufacturing systems in big data environment," Manufacturing Letters, Vol.1, No.1, pp.38-41, 2013.

[5] S. Gallo, T. Murino, and L. Santillo, “Time manufacturing prediction: preprocess model in neuro fuzzy expert system,” Proceeding of The European Symposium on Intelligent Techniques, pp.1-11, 1999.

[6] S. Laschi, M. Fránek, and M. Mascini, "Screen printed electrochemical immunosensors for PCB detection," Electroanalysis, Vol.12, Issue.16, pp.1293-1298, 2000.

[7] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” In Kdd, Vol.96, No.34, 1996.

[8] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," In Proceedings of the ACM-SIAM symposium on Discrete algorithms, pp.1027-1035, 2007.

[9] M. M. Deza and E. Deza, *Encyclopedia of distances*, Springer Berlin Heidelberg, 2009.

[10] C. C. Yang and T. D. Ng, “Analyzing content development and visualizing social interactions in web forum,” IEEE International Conference on Intelligence and Security Informatics, pp.25-30, 2008.

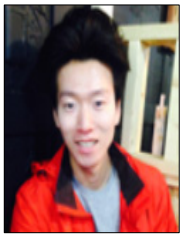
[11] K. E. Krause, “Taxicab geometry,” The

Mathematics Teacher, Vol.66, No.8, pp.695-706, 1973.

[12] G. Chen, S. A. Jaradat, N. Banerjee, T. S. Tanaka, M. S. Ko, and M. Q. Zhang, "Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data," Statistica Sinica, pp.241-262, 2002.

저 자 소 개

최 은 석(Eun-Suk Choi) 정회원



- 2015년 2월 : 배재대학교 경영정보학과(경영학사)
- 2017년 2월 : 충북대학교 컴퓨터과학과(공학학사)

<관심분야> : 빅데이터, 제조 빅데이터, 데이터 수집

김 정 훈(Jeong-Hun Kim) 준회원

- 2017년 2월 : 충북대학교 소프트웨어학과(공학사)
- 2017년 3월 ~ 현재 : 충북대학교 컴퓨터과학(공학석사 재학)

<관심분야> : 데이터베이스, 데이터 클러스터링

아지즈 나스리디노프(Aziz Nasridinov) 정회원

- 2009년 : 동국대학교 컴퓨터공학과(공학석사)
- 2012년 : 동국대학교 컴퓨터공학과(공학박사)
- 2012년 ~ 2014년 : 숙명여자대학교 멀티미디어학과 박사 후 연구원
- 2014년 ~ 2015년 : 동국대학교 컴퓨터공학과 조교수
- 2015년 ~ 현재 : 충북대학교 소프트웨어학과 조교수

<관심분야> : 데이터베이스, 빅 데이터, 데이터마이닝 분산처리 시스템

이 상 현(Sang-Hyun Lee) 정회원

- 2000년 2월 : 한밭대학교 전자계산학과(공학사)
- 2001년 ~ 2002년 : 삼성화학정밀 연구원
- 2002년 ~ 현재 : (주)유라 SW RD 4팀(IT 과) 차장

<관심분야> : 제조 빅데이터, IoT 서비스, ERP, MES

강 정 태(Jeong-Tae Kang) 정회원

- 1987년 2월 : 중앙대학교 기계공학과(공학사)
- 1989년 ~ 1995년 : 삼성전자 연구원
- 1995년 ~ 1999년 : 삼성자동차 연구원
- 1999년 ~ 2003년 : C4Tech
- 2003년 ~ 현재 : (주)유라 IT 총괄 이사

<관심분야> : 제조 빅데이터, IoT 서비스, ERP, MES

류 관 희(Kwan-Hee Yoo) 종신회원

- 1985년 8월 : 전북대학교 전산통계학과(이학사)
- 1988년 2월 : 한국과학기술원 전산학과(공학석사)
- 1995년 8월 : 한국과학기술원 전산학과(공학박사)
- 1988년 1월 ~ 1997년 8월 : 데이콤 선임연구원
- 2003년 7월 ~ 2005년 2월 : 카네기멜론대학교 로보틱스연구소 교환교수
- 1997년 8월 ~ 현재 : 충북대학교 소프트웨어학과 교수

<관심분야> : 컴퓨터 그래픽스, 인공지아보델링, 3D 게임, 3D 콘텐츠