HDBSCAN을 이용한 비지도 공장 설비 부하식별 방법

이준희, 지영민, 권동우*

한국전자기술연구원

{joonhee305, ym.ji, dwkwon*}@keti.re.kr

HDBSCAN-based Unsupervised Load Classification Method for Factory Equipment

Joohee Lee, Youngmin Ji and Dongwoo Kwon*

Korea Electronics Technology Institute (KETI)

요 약

본 논문은 공장 설비의 전류 데이터에 HDBSCAN 클러스터링 알고리즘을 적용하여 실시간으로 설비의 상태를 판단하는 부하식별 방법을 제시한다. 대체로 조업상태의 설비는 비조업 상태의 설비에 비해 전류량이 증가한다. 하지만 모든 설비가 단순히 조업, 비조업 상태와 같은 여러 상태가 존재할 수 있다. 비지도 분류 알고리즘 중 클러스터의 개수를 지정하지 않아도 되는 HDBSCAN 알고리즘을 통해, 사전 공정 정보가 없는 공장 설비에서도 전류 데이터를 통해 여러 상태를 분류하는 연구 결과를 제시한다.

Ⅰ. 서론

스마트 팩토리는 제조혁신을 주도하며 전 세계적으로 확산되고 있다. 스마트 팩토리는 IT를 이용해 모든 제조공정을 통합하고 지능화하여 생산성을 극대화하거나 고객 맞춤형 생산을 구현하는 것이 목적이다.[1] 그러나 이미 구축되어 있는 생산 설비와 통신 및 신호 수집에 대한 호환성 문제 등 복합적인 이유로 기존 설비에 새로운 기능을 위한 인터페이스를 추가하는 것에 어려움이 있다. 또한 장비마다 특성이 다르기 때문에 사전에 장비에 대한 도메인 지식을 갖추어야 한다는 번거로움이 있다.[2]

본 논문에서는 이러한 문제를 개선하고자 설비나 장비에 대한 사전 정보가 없는 상태에서 전류 데이터를 이용한 부하식별 방법을 제시한다. 부하식별이란 공장 설비에서 수집되는 데이터를 통해 설비의 조업, 비조업 등의 상태를 식별하는 방법이다. 전력계 데이터는 전기를 사용하는 장비라면 쉽게 데이터 수집 환경을 구성할 수 있기 때문에 전류 데이터를 사용한다. 전류 데이터를 5분 단위의 패턴으로 나누어 패턴에 대한 클러스터링을 적용해 상태를 분류한다. 이를 통해 전력계 데이터 수집 환경을 통해 사전 정보 없이도 실시간 부하식별이 가능하고, 이를 통한 생산량 예측을 통해 생산성 극대화에 기여할 수 있다.

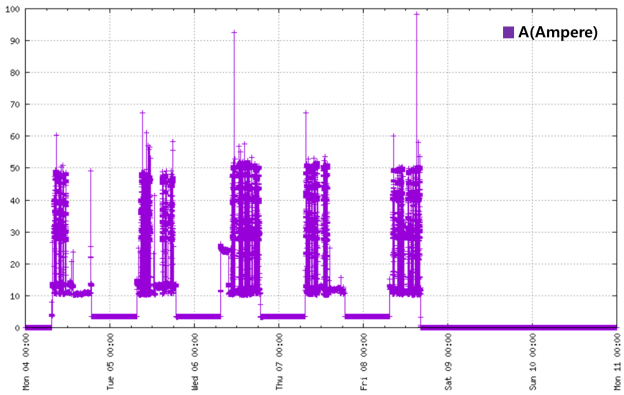
본 논문에서는 HDBSCAN(Hierarchical-DBSCAN, 계층적 밀도 기반 공간 군집) 을 이용해 상태를 분류한다. HDBSCAN은 DBSCAN (Density-based spatial clustering of applications with noise)과 계층적 군집화(Hierarchical clustering)를 결합한 방법으로, 다른 클러스터링 알고리즘에 비해 하이퍼 파라미터가 단순하고, 복잡한 형태의 클러스터도 탐지할 수 있다.[3] 또한 DBSCAN처럼 클러스터에 포함되지 않는 이상치도 분류할 수 있기 때문에 상태의 개수, 주기와 같은 사전 정보가 없는 상황에서 설비의 상태를 분류하기 위한 방법으로 사용했다.

Ⅱ. 본론

부하식별 알고리즘의 단계는 크게 데이터 분석과 전처리, 클러스터링을 통한 라벨 데이터 생성, 실시간 예측을 위한 분류 모델 구축으로 구성된다. 데이터는 실제 운영 중인 공장의 전류 데이터를 사용한다.

2.1 데이터 분석 및 전처리

그림 1은 클러스터링 및 학습에 사용할 데이터로 실시간으로 수집되는 설비의 전류 데이터이다. 대다수의 관제점에서 매 평일의 오전 8시와 오후 4시 사이에 피크가 발생하는 패턴이 반복되는 형태를 나타내고 있다.



(그림 1) 공장 설비 전류 데이터

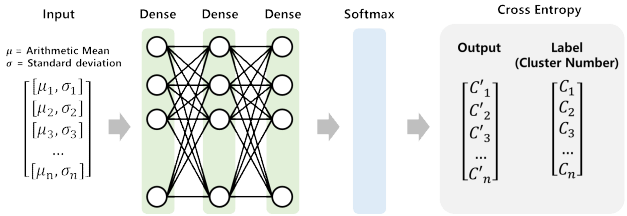
이를 통해 보편적인 근무 시간 대에 설비가 동작하여 전류에 피크가 발생한 것으로 유추할 수 있다. 일주일마다 반복되는 형태고 다양한 상태를 고려하기 위해 본 논문에서는 10초 단위의 2주 치의 데이터를 사용하였다. 결측치는 선형 보간을 통해 처리하였다. 결측치를 처리한 데이터는 클러스터링 및 DNN 학습을 위한 데이터셋으로 변환한다. 5분마다 부하식별을 적용하기 위해 윈도우 크기는 5분으로 설정하고, 슬라이딩 윈도우 기법을 통해 5분 간격(10초 단위로 30개)의 윈도우 크기로 구성한다. HDBSCAN은 유클리디안 거리를 통해 계산하기 때문에 다차원 벡터에 대해서 적절한 거리 계산이 어려울 수 있다. 따라서 5분 단위의 패턴을 각각의 산술평균, 표준 편차를 특성으로 갖는 2차원 벡터로 변환한다.

2.2 클러스터링 적용 및 DNN 학습

클러스터링에 소요되는 시간을 단축하기 위해 병렬 연산이 지원되는 python의 fast-hdbscan 라이브러리를 이용한다. 하이퍼 파라미터인 min_cluster_size를 지정해야 하며, min_cluster_size가 작을수록 작은 사

이즈의 많은 클러스터가 생성되고, min_cluster_size가 클수록 큰 사이즈의 적은 클러스터가 생성된다. 본 논문에서는 전체 데이터셋 개수의 1~10% 사이의 개수를 min_cluster_size로 지정하여 시뮬레이션 한 결과, 경험적으로 다수의 설비에서 적절하게 분류되었다고 판단된 3%를 min_cluster_size로 설정하였다. 클러스터링의 결과값은 클러스터 번호가 할당되지 않은 이상치는 -1, 그 외에는 무작위로 클러스터 번호가 할당된다. 따라서 이상치로 분류된 데이터들을 제거하고, 상태 식별에 활용하기 위해 할당된 클러스터 내 벡터들의 평균값에 따라 클러스터 번호가 오름차순이 되도록 정렬한다.

HDBSCAN은 다수의 데이터에서 군집을 생성하는 방법이므로 계산 복잡도가 높기 때문에 새로운 데이터에 대한 예측을 해야 하는 실시간 부하식별에는 적절하지 않다. 따라서 클러스터링 번호를 라벨데이터로 사용해 DNN을 학습하여 실시간 데이터에 대해 부하식별이 가능하도록 한다. 그림 2는 전처리 한 산술 평균, 표준편차 벡터를 입력받아 클러스터 번호를 예측하는 분류 모델이다. Dense Layer 3개와 분류를 위한 Softmax 레이어로 구성한다. 학습 데이터는 평균, 표준편차 벡터를 사용하며 라벨 데이터로는 클러스터링을 통해 생성된 클러스터 번호를 각 벡터의 라벨로 사용하고, Loss Function은 Cross Entropy를 사용하여 학습한다.

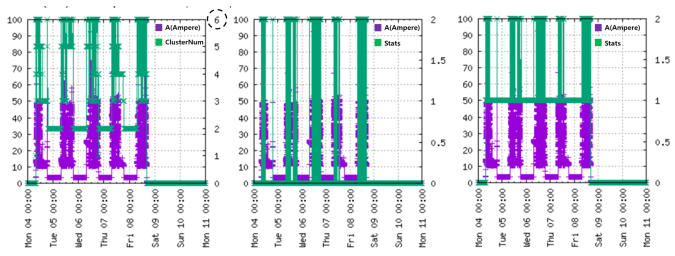


(그림 2) 평균, 표준편차 벡터에 대한 클러스터 번호 예측 모델

2.3 클러스터 번호 응용 및 부하식별 알고리즘 적용 결과

본 논문에서는 Center라는 기준을 선정하여 예측한 클러스터 번호를 조업(work), 비조업(stop), 대기(wait)로 나누는 방법을 제시한다. 조업 상태는 생산 중인 상태를 말하며 비조업 상태는 설비의 전원이 꺼져있거나 생산 중이지 않은 상태, 대기 상태는 조업과 비조업 사이의 상태를 뜻한다. Center는 다음과 같이 정의한다. 각 클러스터에 해당하는 벡터들의 평균으로 정렬한 뒤 x번 클러스터의 백분위 95%에 해당하는 평균값, x-1번 클러스터의 백분위 5%에 해당하는 평균값의 차이가 가장 클 때 x를 Center로 지정한다. Center를 기준으로 3개의 상태로 나눌 때는 0번 클러스터는 비조업, Center 미만은 대기, Center 이상은 조업으로 지정한다. 2개의 상태로 나눌 때는 비조업, 대기 상태를 모두 비조업 상태로 판단한다.

그림 3은 그림 1의 데이터에 부하식별 알고리즘을 적용한 결과이다. 왼쪽부터 cluster 번호, 상태 2개일 때, 상태 3개로 나눌 때이며 Center는 6번 클러스터이다. 상태가 0이면 비조업, 1이면 대기, 2면 조업이다. 전류값이 높은 피크 상태와 전류가 0인 상태, 그 외 상태를 분류하여 각각 조업, 비조업, 대기로 분류했다.



(그림 3) 부하식별 알고리즘 적용 결과

표 1은 실제 가동 중인 설비에서 부하 판단이 가능한 일부 장비의 부하 상태와 부하식별 알고리즘을 적용한 결과를 비교한 정확도 지표이다. 부하 상태 데이터는 한 시간 단위로 장비가 얼마나 켜져 있는지, 얼마나 생산했는지에 대한 밀리초 단위의 데이터로 주어지며, 전원이 꺼져있을 때를 비조업, 생산 중일 때를 조업, 전원은 켜져 있지만 생산은 하지 않는 구간을 대기로 하여 실제 생산 시간과 예측한 시간을 비교한다. 정확도는 (상태가 일치한 시간)/(전체 시간) * 100으로 계산한다.

(표 1) 상태를 2, 3개로 분류했을 때의 정확도(단위 : %)

Stat	A	B	C	D	E	F	G	H	I
2D	91.5	91.6	85.5	90.0	97.4	95.8	95.8	94.9	49.8
3D	14.7	90.7	59.8	89.8	97.1	24.6	95.7	73.0	39.1

2D는 상태를 조업, 비조업 2가지로 나타낸 경우, 3D는 상태를 조업, 대기, 비조업 3개로 나타낸 경우이다. 2D의 경우 대부분의 관제점에서 90% 이상의 정확도로 조업과 비조업 상태를 식별하는 것을 확인할 수 있다. 3D의 경우 일부 관제점에서 클러스터 단계에서 대기 상태를 분류하지 못한 케이스가 존재하여 상대적으로 정확도가 낮게 나타났다. 백분위 5%, 95% 값으로 Center를 결정할 때 극단값이 포함될 경우 적절한 Center를 결정하지 못하여 정확도가 낮아진 것으로 판단된다. 따라서 거리를 계산할 때 사용하는 값을 백분위 15%, 85% 정도로 확장하는 등 Center를 결정하는 방법을 개선하여 해결하고자 한다.

Ⅲ. 결론

본 논문에서는 HDBSCAN을 이용해 전류 데이터의 패턴을 통해 부하를 식별하는 방법을 적용하고 결과를 분석했다. 실제 조업 여부 라벨데이터가 없더라도 전류 데이터를 통해 상태를 분류할 수 있었다. 클러스터링을 통해 전류 데이터의 패턴을 분류하고 Center를 선정해 해당 클러스터들을 조업, 비조업, 대기 상태로 나눴다. 실제 조업 여부 데이터와 비교하여 정확도를 측정했을 때, 실제 조업 중인 상태를 잘 식별한 것으로 판단된다. 하지만 대기 상태는 상대적으로 분류 정확도가 떨어졌다. 또한 공장의 특성상 조업 상태가 비조업 상태보다 더 많기 때문에 클래스 불균형의 문제도 존재한다. 추후 연구에서는 Center 결정 방법에 대한 보강을 통해 대기 상태의 분류 정확도를 높이고, 클래스 불균형 문제를 해소하기 위해 데이터 증강 기법을 적용하고자 한다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. 20202020900290)

참 고 문 헌

[1] Y. Kim and B. Yang, "Extracting Urban Areas of Interest Using HDBSCAN Clustering Method", Journal of the Korean Cartographic Association, vol. 23, no. 1, pp. 67-77, Apr 2023. (https://doi.org/10.16879/jkca.2023.23.1.067)

[2] D. Lee. "IIoT Architecture and Method of Collecting Signal for Build Smart-Factory Environment" M.S. thesis, Graduate School, Gachon University, Gyeonggi-do, Korea, Feb 2022.

[3] H. Yang, "Policy Measures for Revitalizing the Artificial Intelligence-Based Smart Factory", The Journal of Korean Institute of Communications and Information Sciences, vol. 45, no. 9, pp. 1659-1665, Sep. 2020. (https://doi.org/10.7840/kics.2020.45.9.1659)