KIPS Tr. Software and Data Eng.
Vol.6, No.10 pp.465~472 pISSN: 2287-5905



https://doi.org/10.3745/KTSDE.2017.6.10.465

# A Novel of Data Clustering Architecture for Outlier Detection to Electric Power Data Analysis

Se Hoon Jung† · Chang Sun Shin†† · Young Yun Cho†† · Jang Woo Park†† · Myung Hye Park††† · Young Hyun Kim†††† · Seung Bae Lee††††† · Chun Bo Sim††

## ABSTRACT

In the past, researchers mainly used the supervised learning technique of machine learning to analyze power data and investigated the identification of patterns through the data mining technique. Data analysis research, however, faces its limitations with the old data classification and analysis techniques today when the size of electric power data has increased with the possible real-time provision of data. This study thus set out to propose a clustering architecture to analyze large-sized electric power data. The clustering process proposed in the study supplements the K-means algorithm, an unsupervised learning technique, for its problems and is capable of automating the entire process from the collection of electric power data to their analysis. In the present study, power data were categorized and analyzed in total three levels, which include the row data level, clustering level, and user interface level. In addition, the investigator identified K, the ideal number of clusters, based on principal component analysis and normal distribution and proposed an altered K-means algorithm to reduce data that would be categorized as ideal points in order to increase the efficiency of clustering.

Keywords : Data Analysis, Electric Power, Unsupervised Learning, Outlier, PCA

# 전력데이터 분석에서 이상점 추출을 위한 데이터 클러스터링 아키텍처에 관한 연구

정 세 훈† · 신 창 선†† · 조 용 윤†† · 박 장 우†† · 박 명 혜††† · 김 영 현†††† · 이 승 배††††† · 심 춘 보††

## 요 약

과거에는 전력데이터를 분석하는 기법으로 주로 기계학습의 지도학습 기법을 많이 활용하였고 데이터 마이닝 기법을 통한 패턴 검출을 주로 연구하였다. 그러나 전력데이터의 규모 커지고 실시간 데이터 공급이 가능해진 현재에는 과거의 데이터 분류 및 분석 기법을 통한 데이터 분석 연구는 한계가 존재한다. 이에 본 논문에서는 큰 규모의 전력데이터를 분석할 수 있는 클러스터링 아키텍처를 제안한다. 제안하는 클러스터링 프로세스는 비지도학습기법인 K-means 알고리즘의 문제점을 보완하고 전력데이터 수집과 분석까지의 모든 과정을 자동화할 수 있는 프로세스이다. 총 3 Level로 구분하여 Row Data Level, Clustering Level, User Interface Level로 구분하여 전력데이터를 분류 및 분석한다. 또한 클러스터링의 효율성 향상을 위하여 주성분분석 및 정규분포기반의 최적의 클러스터 수 K값 추출과 이상점으로 분류되는 데이터 감소를 위한 변형된 K-means 알고리즘을 제시한다.

키워드 : 데이터 분석, 전력데이터, 비지도학습, 이상점, 주성분분석

## 1. 서 론

기존의 전력 분석(Electric Power) 연구는 주로 지도학습 기법(Supervised Learning Method) 중 연관분석(Association Analysis), 상관분석(Correlation Analysis), 회귀분석(Regression Analysis)과 또는 시계열 분석(Time Series Analysis)과 같이 연속형 전력 데이터와 범주형 전력 데이터를 1차적으로

※ 이 논문은 한국전력공사의 재원으로 2015년 선정된 자체 연구개발 과제의 지원을 받아 수행된 것임(과제번호: R15CA03).
† 준 회 원 : 광양만권 SW융합연구소 팀장
†† 정 회 원 : 순천대학교 정보통신·멀티미디어공학부 교수
††† 정 회 원 : 한전 전력연구원 책임연구원
†††† 비 회 원 : 한전 전력연구원 선임연구원
††††† 비 회 원 : 한전 전력연구원 책임연구원
Manuscript Received : June 12, 2017
Accepted : July 4, 2017
* Corresponding Author : Chun Bo Sim(cbsim@sunchon.ac.kr)

분석하고 분석된 자료를 기초로 학습데이터(Learning Data)를 생성한 후 새로 입력되는 데이터를 학습하는 기법이 주로 연구 되고 있다[1-4].

그러나 시대의 변화에 따른 전력 사용 증가 및 전력 사용 증가로 인해 발생되는 전력 센서노드 데이터의 형태 변화, 비정형 전력 분석 데이터가 늘어나면서 기존의 연구 방식으로 처리하는 것에는 몇 가지 문제점과 추가적인 해결 방법이 필요하다[5]. 새롭게 생성되는 전력데이터는 기존의 학습 모델로 분석하기 보다는 실시간적으로 분류 및 분석할 수 있는 모델이 필요한 실정이다[6-7]. 실시간 분석 모델은 기본적으로 레이블이 존재하지 않는 새롭게 생성된 데이터에 대하여 전처리 과정인 클러스터링 단계가 선행되어야 한다. 클러스터링 기법은 1960년대부터 연구되어진 분야이며, 지금까지 문제점을 보완하면 꾸준히 연구되고 있다. 특히 기본적인 K-means 알고리즘은 빅데이터에 적용하여 클러스터링을 진행하는데 있어 데이터 처리 구조, 처리 용량, 처리 비용에서 많은 문제점이 발생되고 있다[8]. K-means 알고리즘의 문제점은 다음과 같다. 첫 번째로는 클러스터링을 하기 위한 클러스터 수 K값을 임의적인 선택을 하거나 또는 휴리스틱한 방법을 통해 최적의 해를 찾아야 하는 문제점이 있어 클러스터링의 비용이 증가하는 결과를 초래한다. K값 선택의 문제점은 빅데이터 이전 시기에 생성된 데이터의 크기와 규모에서 임의적인 선택이 가능하였지만, 메가 데이터 또는 빅데이터 시대에서 발생되는 전력 로우 데이터는 임의적인 선택이 아닌 최적의 K값을 자동으로 찾아주는 기법이 필요하다. 두 번째로는 클러스터링 진행 시 이상점(Outlier) 추출로 인한 비용 증가 문제이다. 클러스터링 기법은 크게 4가지로 구분된다. 독점 클러스터링, 중복 클러스터링, 계층 클러스터링, 확률 클러스터링으로 분류된다. 중복 클러스터링을 제외한 3개의 클러스터링은 모호한 클러스터 경계면이 존재할 수 있으며, 클러스터에 인접해 있지 않은 데이터는 이상점으로 쉽게 분류한다. 이는 클러스터의 효율성과 정확률 측면에서 큰 문제점으로 지적되고 있다. 특히, 전력데이터에서는 기본 범위에서 움직이는 센서 데이터가 클러스터 내부에서 강한 밀집도를 가지고 있기 때문에 클러스터 외부에 존재하는 센서 데이터를 이상점으로 인식하는 경우가 종종 발생하고 있다.

이에 본 논문에서는 기존 K-means 알고리즘의 문제점을 분석하고 자동화된 클러스터 K값 선택과 이상점 해결을 기반한 전력데이터 클러스터링 아키텍처를 제안한다. 이를 위해 전력데이터 클러스터링 아키텍처는 기본적으로 K-means 알고리즘의 문제점을 보완하는 새로운 알고리즘을 포함한다. 데이터 클러스터링 아키텍처는 크게 3개의 레벨로 구분하여 설계한다. 전력데이터 수집과 전처리 과정이 포함된 Row Data Level과 변형된 K-means 알고리즘이 적용되는 Clustering Level, 전력데이터 분석과 결과를 확인하는 User Interface Level로 구분하여 설계를 진행한다. Row Data Level에서는 데이터 수집과 불필요한 데이터 제거 및 데이터 정규화과정이 포함된다. Clustering Level에서는 본 연구에서 제안하는 변형된 K-means 알고리즘이 포함된다. 변형된 알고리즘은 크게 두 가지 측면에서 알고리즘을 제안한다. 자동화된 K값 선택을 위하여 클러스터링을 위한 데이터의 다차원 데이터를 분석하여 전체 데이터를 설명할 수 있는 기준점을 확보하는 주성분을 최적화된 클러스터 수 K값으로 선택한다. 또한 이상점 문제를 해결하기 위하여 초기 클러스터의 후보 중심점을 정규분포에 따라 95% 이상에 포함된 데이터를 선택하고 유클리디안 측정법에 따라 초기 클러스터의 중심점으로 선택한다. 이를 통해 이상점으로 처리될 수 있는 데이터를 미리 확보하여 이상점의 발생 빈도를 낮추고 클러스터의 효율성을 높인다. 마지막으로 User Interface Level은 시스템 사용자의 인터페이스를 담당하는 레벨로 C#과 Python을 이용한 전력데이터 분석 인터페이스와 전력데이터의 가설 및 검증 단계가 포함된다.

## 2. 관련 연구

### 2.1 K-means

K-means[8-9] 알고리즘은 군집의 수 K를 미리 정하여 각 데이터가 특정 클러스터에 배치되는지를 분석하는 기법이며, 재배치 방법을 활용한 비 계층적 클러스터링 알고리즘으로 계층적 알고리즘보다 연산 속도가 빠르고 간결하여 클러스터링 알고리즘에서 대표적으로 활용되는 기법이다. K-means 알고리즘의 절차는 다음과 같다.

① 사용자에 의해 미리 클러스터 수(K)를 결정한다.
② 선정된 클러스터 수 K개의 각 클러스터에 전체 데이터 중 한 개를 각 클러스터에 포함시킨다.
③ 클러스터링을 위한 모든 데이터는 거리 계산법을 통해 가장 가까운 클러스터의 중심으로 배속된다.
④ ③ 단계 과정을 실시한 후 배속된 데이터들의 중심을 해당 클러스터의 새로운 중심점으로 정하고 클러스터를 재 할당한다.
⑤ 데이터의 이동이 없을 때까지 ③단계와 ④단계를 반복적으로 실시한다.

지도학습이 매 입력 벡터가 입력될 때마다 가중 벡터를 갱신하는 반면 K-means 알고리즘은 모든 입력 벡터들이 입력된 후 동시에 가중 벡터들을 갱신한다. 군집화를 분류하는 기준은 각 클러스터간 거리 및 클러스터간 비유사도(Dissimilarity)와 같은 비용 함수(Cost Function)을 최소화하는 기법이다. 같은 클러스터내 데이터 오브젝트끼리의 유사도는 증가한다. 다른 클러스터에 있는 데이터 오브젝트와의 유사도는 감소한다. K-means 알고리즘은 각 클러스터의 중심(Centroid)과 클러스터내의 데이터 오브젝트와 거리의 제곱합을 비용 함수로 정하고, 비용 함수값을 최소화하는 방향으로 각 데이터 오브젝트의 클러스터분류를 반복해 줌으로써 군집화를 수행하게 된다. 클러스터 내부 거리(IntraCD : Intra Cluster Distance)는 각 클러스터의 중심에

서 해당 클러스터에 할당된 모든 입력 벡터들까지 거리를 더한 값 이다. 클러스터 사이 거리(ICD : Inter Cluster Distance)는 두 클러스터의 가중 벡터 사이 거리이다. 오차 Error는 Equation (1)과 같이 모든 클러스터에 대한 클러스터 내부 거리의 합을 더하고 모든 클러스터 쌍에 대한 클러스터 사이 거리의 합을 빼서 계산한다. $\beta$와 $\gamma$는 가중치이다.

$$Error = \beta \sum_{i=0}^{k} (IntraCD) - \gamma \sum_{i=0}^{k} (ICD) \quad (1)$$

### 2.2 Principal Component Analysis

주성분분석[10]은 데이터의 다차원 입력 벡터를 가능한 정보 손실을 낮춘 후 낮은 차원의 벡터로 환원시키는 비지도학습 기법이다. 몇 개의 주성분 값으로 나타내는 다변량 데이터 처리 기법 중의 하나이다. n차원의 벡터가 있을 경우, Equation (2)와 Equation (3)을 적용해 나온 평균 벡터와 분산 공분산 행렬을 통해 고유벡터를 구한 뒤에 대응되는 고유값 크기에 따라 고유벡터를 정렬하여 새로운 행렬을 추가한다.

$$m_x = \frac{1}{M} \sum_{k=1}^{M} x_k \quad (2)$$

$$C_x = \frac{1}{M} \sum_{k=1}^{M} x_k x_k^T - m_k m_k^T \quad (3)$$

이 새로운 행렬은 변환 행렬로 적용해 Equation (4)와 같이 벡터 x를 벡터 y로 변환하면, y열에 있는 새 변수들은 비상관성을 가지며 단조 감소 분산 순서로 배열되어 분산 값이 큰 주성분들로 차원을 줄일 수 있다.

$$y = nMatrix(x - m_x) \quad (4)$$

## 3. 전력데이터 클러스터링 아키텍처 설계

### 3.1 전력데이터 클러스터링 아키텍처 구성

본 연구에서는 전력데이터 분석을 위하여 클러스터링 아키텍처를 제안한다.

클러스터링 아키텍처는 변형된 K-means 알고리즘을 기반으로 Row Data Level, Clustering Level, User Interface Level의 3단계로 구분하여 설계한다. Fig. 1은 본 연구에서 제안하는 전력데이터 클러스터링 아키텍처이다.

### 3.2 Row Data Level

본 연구에서는 전력데이터 분석을 위하여 Row Data Level에서는 데이터 수집과 불필요한 데이터 제거, 데이터 정규화 단계가 포함되어 있다. 수집되는 센서노드 데이터는 전주번호, 장비위치, 온도, 습도, 피치, 롤, 조도, 자외선, 압력, 배터리 잔량, 주기를 확인할 수 있다. 수집된 Row Data는 불필요 데이터 제거를 통해 특정 Column만 추출하여 데이터 정규화 과정을 거치게 된다. 데이터 정규화 과정에서는 데이터 범위를 가용범위 이내로 수치화하여 분석 모델에 적용할 수치 범위로 재가공하게 된다.

### 3.3 Clustering Level

Clustering Level에서는 변형된 K-means 알고리즘적용하며, 크게 두 가지 측면에서 알고리즘을 제안한다. 자동화된 K값 선택을 위하여 클러스터링을 위한 데이터의 다차원 데이터를 분석하여 전체 데이터를 설명할 수 있는 기준점을 확보하는 주성분을 최적화된 클러스터 수 K값으로 선택한다. 또한 이상점 문제를 해결하기 위하여 초기 클러스터의 후보 중심점을 정규분포에 따라 95% 이상에 포함된 데이터를 선택하고 유클리디안 측정법에 따라 초기 클러스터의 중심점으로 선택한다. 이를 통해 이상점으로 처리될 수 있는 데이터를 미리 확보하여 이상점의 발생 빈도를 낮추고 클러스터의 효율성을 높인다.

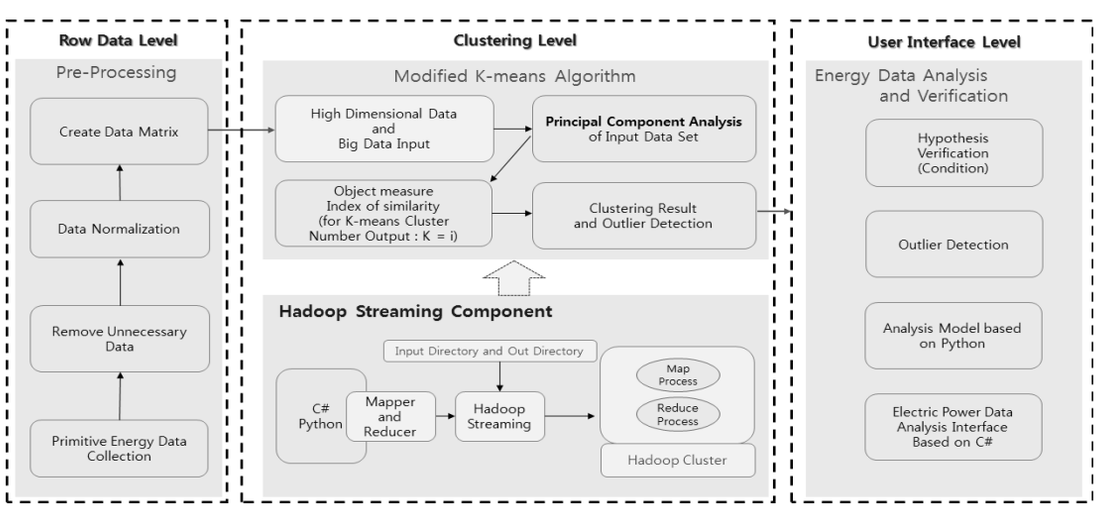



Fig. 1. Electric Power Data Clustering

### 3.4 User Interface Level

시스템 사용자의 인터페이스를 담당하는 Level로 C#과 Python을 이용한 전력데이터 분석 인터페이스와 분석 모델을 설계한다. 또한 전력데이터는 지도학습과 비지도학습 모두 전주의 센서노드 이상 유무 및 숨겨진 패턴을 분석하기 위하여 전력데이터 분석의 가설 및 검증 단계가 포함되어야 한다. 전력데이터 분석 시스템의 인터페이스는 관리자 인터페이스만 정의되며, 정규화과정을 거친 수치데이터를 기준으로 2차원형태의 결과 인터페이스를 제공한다.

### 3.5 최적의 클러스터 수 K값 선정 알고리즘

주성분 분석(Principal Component Analysis)의 공분산(Covariance)은 주어진 전력데이터에서 선택된 클러스터와 다른 클러스터간의 연관성을 확인할 수 있다. 이를 바탕으로 클러스터 경계는 주성분 분석을 통해 구분할 수 있다. 클러스터 지표 벡터의 주성분은 $v_i$로 정의되며, Equation (5)와 같이 정의한다. 초기 중심 값의 최적화된 조건에 만족하는 범위는 Equation (6)과 같이 정의된다.

$$k_1 = \{i | v_1(i) \le 0\}, k_2 = \{i | v_1(i) > 0\} \tag{5}$$

$$n\overline{y^2} - \lambda_1 < k_{k=2} < n\overline{y^2} \tag{6}$$

임의 클러스터 내의 데이터 데이터와 임의로 선택된 중심점과의 평균 거리 측정은 유클리드 거리 제곱의 합으로 정의한다. Equation (7)은 임의로 선택된 중심점과 입력된 모든 데이터 데이터와의 측정한 평균 거리이다. Equation (9)의 $S(k)$는 Equation (7)의 임의 클러스터 내의 데이터에 대한 평균 거리인 $A_k$(응집도)와 Equation (8)의 임의 클러스터 외의 데이터에 대한 평균 거리의 최소값인 $B_k$(분리도)의 차에 의한 결합으로 최대값을 주는 클러스터 최적 비유사도 $S(k)$가 클러스터 수 K가 되며, Equation (10)과 같다. 단, $S(k)$는 -1과 1사이의 값으로 1에 가까울수록 최적화된 클러스터 수로 선택한다.

$$A_k = \sum_{k=1}^{k} \sum_{i \in C_k} (X_i - m_k)^2 \tag{7}$$

$$B_k = \min\{(X_i - m_k)^2\} \tag{8}$$

$$S(k) = \frac{1}{N} \sum_{i=1}^{k} \frac{B_k - A_k}{\max\{A_k, B_k\}} \tag{9}$$

$$S(k) = \begin{cases} 1 - A_k/B_k, \text{ if } A_k < B_k \\ A_k/B_k - 1, \text{ if } A_k > B_k \end{cases} \tag{10}$$

$$-1 \le S(k) \le 1$$

### 3.6 초기 이상점 선택을 위한 클러스터 중심점 알고리즘

초기 중심점을 선택하기 위하여 데이터간의 유사도 및 밀집도의 분포와 표준 정규 분포의 위치를 활용한다. 입력되는 분석 전력데이터 중 첫 번째 클러스터 $C_1$의 중심점 $m_k$는 표준 정규 분포 $\phi_{\mu,\sigma^2}(x_k y_k)$를 기준으로 $P(\overline{X} \ge x_k y_k) = \pm 0.9$ 이상 범위에 분포할 경우 초기 중심점에 대한 관측값 항목으로 분류한다. 표준 정규 분포를 따라서 초기 클러스터 중심점이 될 관측값은 $k^2$의 이하이며, 해당 조건은 입력 데이터 집합 $X$, 평균 $\mu$, 표준편차 $\sigma$가 존재한다. 클러스터 수 K가 0개 이상의 클러스터링을 진행할 경우 Equation (11)에 만족한다.

$$P(|x-\mu| \ge k\sigma) \le k^2 \tag{11}$$

$$P(|x-\mu| \ge k\sigma) = P((x-\mu)^2 \ge k^2\sigma^2)$$

$$= \frac{k^2\sigma^2}{E[(x-\mu)^2]}$$

$$= \frac{k^2\sigma^2}{\sigma^2}$$

$$= k^2$$

또한 중심점의 관측값 데이터에서 1개 이상의 데이터일 경우 데이터 간 유사도 및 밀집도 측정하기 위한 데이터 거리 $A_k$의 측정식은 Equation (12)와 같이 정의한다. 각 데이터간 거리 측정에 대한 평균값 $AVG_k$은 Equation (13)과 같이 정의한다.

$$A_k = d(x_{ki}, x_{li}) = \sum_{k=1}^{k} \sum_{i \in C_1} (X_i - X_k)^2 \tag{12}$$

$$AVG_k = \frac{1}{k} \sum_{k=1}^{k} \sum_{i \in C_1} (X_i - X_k)^2 \tag{13}$$

첫 번째 클러스터 $C_1$의 중심점인 $m_1$을 기준으로 중심점을 제외한 모든 데이터 $x_i$에 대하여 데이터간 거리 $A_k$를 측정한다. 측정 결과가 값이 최대가 되는 데이터를 두 번째 클러스터 $C_2$의 중심점인 $m_2$에 배속되는 것으로 정의하며, Equation (14)와 같이 두 번째 클러스터 배속에 대한 식을 정의한다.

$$C_2(m_2) = a_i \leftarrow {}_{1 \le i \le n}^{\max}(A_k) \tag{14}$$

$$= {}_{1 \le i \le n}^{\max} \| x_i - m_1 \|$$

$$= \| a_i - m_1 \|$$

## 4. 전력데이터 클러스터링 프로세스 모델 적용

### 4.1 실험 데이터 구성

본 연구에서는 전력데이터 클러스터링 프로세스 모델 적용과 성능평가를 위하여 2016년 4월 특정지역 전주의 센서노드를 기반으로 제안하는 프로세스 모델의 실험을 진행한다. 전주 센서노드는 변압기 본체, 부하 개폐기 완금, 전주, 통신용함체로 구분되어 있으며, 각각 전주번호, 장비위치, 온

도, 습도, 피치, 롤, 조도, 자외선, 압력, 배터리 잔량, 주기로 구분하여 데이터를 수집한다. 그리고 전주 기본정보인 위치, 코드, 시설, 일시, 날짜, 시간, Pole, Position 정보를 수집하여 분석 시스템에 적용한다. 총 실험데이터는 4곳의(A: 6,318개, B: 3,881개, C: 7,540, D: 7,239)에서 총 24,978개의 데이터를 활용하였다.

### 4.2 전력데이터 클러스터링 프로세스 인터페이스

Fig. 2는 C# 기반의 전력데이터 클러스터링 프로세스 인터페이스이며, 주성분분석을 통해 자동 선택된 클러스터 수에 따라 클러스터링을 진행한 결과 인터페이스이다. B 전주의 전주장비에 대한 Temp.-Pitch 항목에 대한 클러스터링 결과 UI이다.

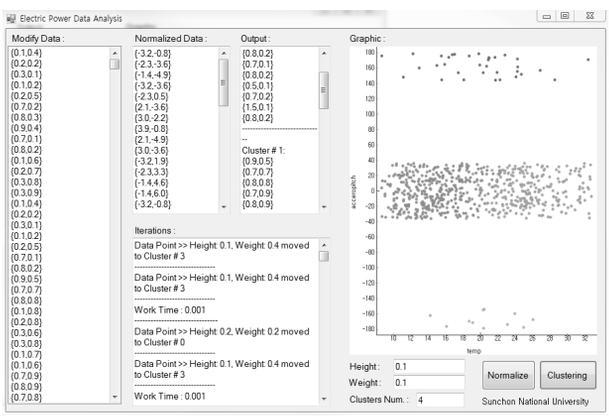
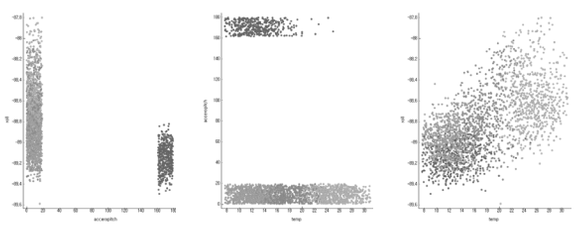



Fig. 2. Clustering Result Interface

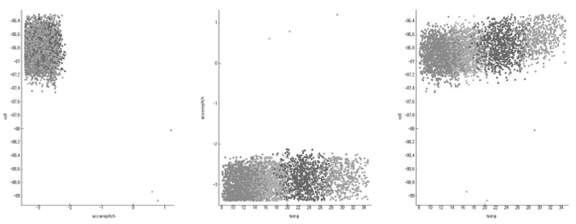
### 4.3 전력데이터 클러스터링 프로세스 이상점 추출 결과

본 연구에서는 전주의 전력데이터를 제안하는 변형된 K-means 클러스터링 프로세스 모델에 적용하여 자동화된 K값 선정과 이상점 추출에 관하여 결과를 확인한다. 프로세스 모델 적용을 위해서 4곳의 지역에 완금과 전주를 구분하여 온도, 피치, 롤간의 클러스터를 진행하였고 최적의 클러스터 수 K와 이상점의 발생 규모의 비교평가를 진행한다. 주성분분석을 통해 다차원의 정보는 4차원으로 줄일 경우 전체 데이터의 80% 설명할 수 있는 결과를 도출하여, 최적의 클러스터 수 k값은 4로 자동 설정되어 각 지역의 클러스터링을 진행하였다. Fig. 3은 A의 전주에 대한 클러스터링 결과이며, Fig. 4는 A의 완금에 대한 클러스터링 결과이다.



Pitch-Roll Temp-Pitch Temp-Roll

Fig. 3. Clustering Result by K=4 (Position=0)

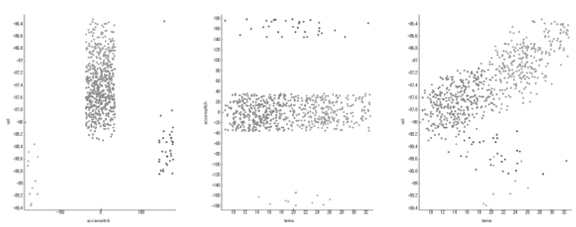


Pitch-Roll Temp-Pitch Temp-Roll

Fig. 4. Clustering Result by K=4 (Position=2)

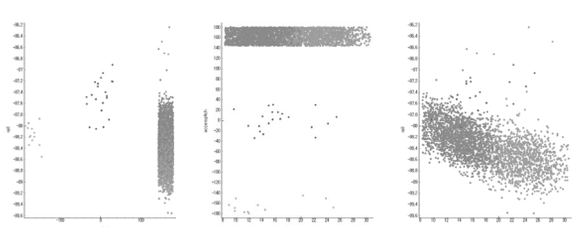
최종 이상점 검출은 전주일 경우 23개, 완금일 경우 43개로 총 63개의 이상점이 검출되었다.

Fig. 5는 B의 전주에 대한 클러스터링 결과이며, Fig. 6은 B의 완금에 대한 클러스터링 결과이다. 최종 이상점 검출은 전주일 경우 16개, 완금일 경우 34개로 총 50개의 이상점이 검출되었다.



Pitch-Roll Temp-Pitch Temp-Roll

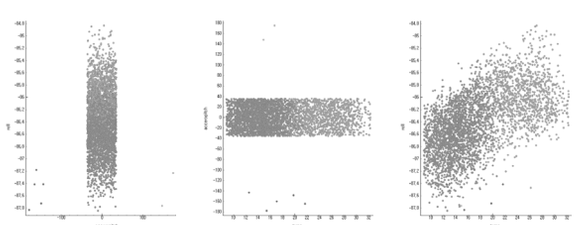
Fig. 5. Clustering Result by K=4 (Position=0)



Pitch-Roll Temp-Pitch Temp-Roll

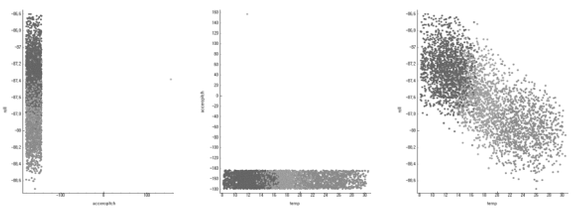
Fig. 6. Clustering Result by K=4 (Position=2)

Fig. 7은 C의 전주에 대한 클러스터링 결과이며, Fig. 8은 C의 완금에 대한 클러스터링 결과이다. 최종 이상점 검출은 전주일 경우 35개, 완금일 경우 41개로 총 76개의 이상점이 검출되었다.



Pitch-Roll Temp-Pitch Temp-Roll

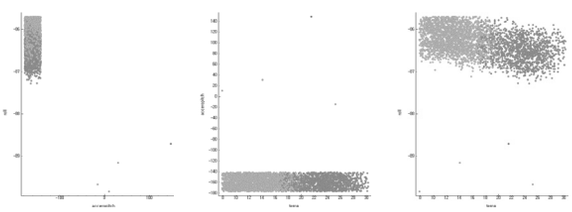
Fig. 7. Clustering Result by K=4 (Position=0)



Pitch-Roll Temp-Pitch Temp-Roll

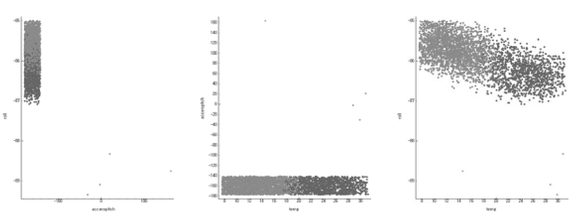
Fig. 8. Clustering Result by K=4 (Position=2)

Fig. 9는 D의 전주에 대한 클러스터링 결과이며, Fig. 10은 D의 완금에 대한 클러스터링 결과이다. 최종 이상점 검출은 전주일 경우 44개, 완금일 경우 35개로 총 79개의 이상점이 검출되었다.



Pitch-Roll Temp-Pitch Temp-Roll

Fig. 9. Clustering Result by K=4 (Position=0)



Pitch-Roll Temp-Pitch Temp-Roll

Fig. 10. Clustering Result by K=4 (Position=2)

### 4.4 이상점 추출에 따른 전력데이터 분석

본 연구에서는 K-means 알고리즘의 클러스터 분석을 위하여 이상점 초기 중심점 접근 방식을 통하여 이상점에 대한 민감도를 비교평가 하였다. 제안하는 알고리즘을 적용하지 않을 경우 발생되는 이상점 개수와 제안하는 알고리즘을 적용하였을 경우 발생되는 이상점의 개수는 Table 1과 같다. 4곳 모두 제안하는 알고리즘으로 적용할 경우 이상점 검출 데이터가 줄어들었다. A는 32개, B는 24개, C는 15개, D는 18개의 데이터가 이상점으로 검출되지 않은 결과를 보였다. 이러한 결과는 초기 중심점을 정규 분포도를 기준으로 최외각 지점 부분부터 클러스터의 초기 중심점으로 지정하기 때문에 결과적으로 이상점의 개수가 전체적으로 줄어들었다. 또한 Table 2를 통해 4곳의 전력데이터에서 전체 데이터를 기준으로 2.0.1%~3.45%까지의 이상점 검출이 반영된다. 또한 B의 전주 이상점 추출 비율이 상대적으로 높게 나타난 결과는 피치와 롤의 변화값이 상대적으로 크지는 않지만 작은 범위 내에서 상하좌우로 움직이는 값들에 대하여 하나의 클러스터로 인식하지 못하면서 상대적으로 이상점의 비율이 높게 나타나고 있다. 또한 피치 항목과 결합한 클러스터링 진행 시 주 분포 구역을 벗어난 경우 새로운 클러스터로 인식하여 상대적으로 이상점의 추출이 낮게 측정되었다.

Table 1. Comparison of Inlier-Outlier Detection (K=4)

| Part | Position | lier | Non-Proposed | Proposed |
|---|---|---|---|---|
| A | 0 | Inlier | 2,419 | 2,429 |
| | | Outlier | 33 | 23 |
| | 2 | Inlier | 3,801 | 3,823 |
| | | Outlier | 65 | 43 |
| B | 0 | Inlier | 648 | 654 |
| | | Outlier | 22 | 16 |
| | 2 | Inlier | 3,159 | 3,177 |
| | | Outlier | 52 | 34 |
| C | 0 | Inlier | 3,688 | 3,701 |
| | | Outlier | 48 | 35 |
| | 2 | Inlier | 3,761 | 3,763 |
| | | Outlier | 43 | 41 |
| D | 0 | Inlier | 3,633 | 3,644 |
| | | Outlier | 55 | 44 |
| | 2 | Inlier | 3,509 | 3,516 |
| | | Outlier | 42 | 35 |

Table 2. Inlier-Outlier Detection Ratio(k=4)

| Part | Position | lier | Ratio |
|---|---|---|---|
| A | 0 | Inlier | 99.062% |
| | | Outlier | 0.938% |
| | 2 | Inlier | 98.888% |
| | | Outlier | 1.112% |
| B | 0 | Inlier | 97.612% |
| | | Outlier | 2.388% |
| | 2 | Inlier | 94.941% |
| | | Outlier | 1.058% |
| C | 0 | Inlier | 99.063% |
| | | Outlier | 0.937% |
| | 2 | Inlier | 98.822% |
| | | Outlier | 1.077% |
| D | 0 | Inlier | 98.807% |
| | | Outlier | 1.193% |
| | 2 | Inlier | 99.014% |
| | | Outlier | 0.985% |

## 5. 결 론

기존 K-means 알고리즘을 통해 전력데이터를 분류 및 분석할 경우 클러스터 K값 선택과 이상점 추출에 대하여 성능적인 향상을 보여주지 못했다. 이에 본 연구에서는 전력데이터 분석을 위한 자동화된 클러스터 K값과 이상점 추출을 위하여 K-means, 정규분포, 주성분분석기반의 데이터

클러스터링 아키텍처를 제안하였다. 성능평가를 위한 실험 데이터는 2016년 4월 특정 지역 전신주의 센서노드에서 수집된 정보를 활용하였으며, 전신주의 전주와 완금의 온도, 피치, 롤 데이터를 분류 및 분석하였다. 입력되는 전력데이터의 다차원 정보를 분석하여 저차원 수준이고 전체 데이터를 포괄적으로 설명할 수 있는 80%의 기준 차원을 추출하고 추출된 차원에서 최적의 K값을 인식하도록 하였다. 또한 기계학습에서 발생할 수 있는 이상점의 형성 조건을 분석하고 분석된 결과를 바탕으로 초기 접근을 이상점으로 지정하여 이상점의 최종 추출 개수를 줄이고자 하였다. 성능평가 결과 전력데이터 4개의 24,978개의 데이터를 학습한 결과 제안하는 알고리즘과 기존 K-means 알고리즘의 차이를 확인할 수 있었다. 총 4개의 전주에서 완금과 전주를 각각 구분하여 성능평가를 진행하였고 총 89개 이상점 감소와 전체 데이터 대비 평균 0.5% 감소율을 확인하였다. 그러나 일부 데이터에 한하여 클러스터의 경계면과 외곽지점에서 새로운 클러스터를 형성하면서 다른 데이터에 영향을 주는 부분은 추후 연구를 통해 개선해야 할 부분이다.

## References


[1] E. Y. Hong and M, Y, Park, "Unsupervised Learning Model for Fault Prediction Using Representative Clustering Algorithms," *Journal of Software and Data Eng.*, Vol.3, No.2, pp.57-64, 2014.

[2] J. M. Lee, J. Lee, and J. S. Kim, "Ontology-based Monitoring Approach for Efficient Power Management in Datacenters," *Journal of Korean Institute of Information Scientists and Engineers*, Vol.42, No.5, pp.580-590, 2015.

[3] D. I. Park and S. H. Yoon, "Clustering and classification to characterize daily electricity demand," *Journal of the Korean Data & Information Science Society*, Vol.28, No.2, pp.395-406, 2017.

[4] J. H. Park, H. G. Lee, J. H. Shin, and K. H. Ryu, "Analysis and Prediction of Power Consumption Pattern Using Spatiotemporal Data Mining Techniques in GIS-AMR System," *Journal of Information Processing Systems*, Vol.16, No.3, pp.307-316, 2009.

[5] S. H. Yoon and Y. J. Choi, "Functional clustering for electricity demand data: A case study," *Journal of the Korean Data & Information Science Society*, Vol.26, No.4, pp.885-894, 2015.

[6] M. H. Park, Y. H. Kim, and S. B. Lee, "A study on the Development of Energy IoT Platform," *KIPS Tr. Comp. and Comm. Sys.*, Vol.5, No.4, pp.311-318, 2016.

[7] S. H. Ryu, H. S. Kim, D. E. Oh, and J. K. No, "Customer Load Pattern Analysis using Clustering Techniques," *KEPCO Journal on Electric Power and Energy*, Vol.2, No.1, pp.61-69, 2016.

[8] S. H. Jung, "A Novel on Hybrid Machine Learning Method based on Big Data Mining," Doctor Thesis, Sunchon National University, 2017.

[9] K. Zhang, W. Bi, X. Zhang, X. Fu, K. Zhou, and L. Zhu, "A New K-means Clustering Algorithm for Point Cloud," *International Journal of Hybrid Information Technology*, Vol.8, No.9, pp.157-170, 2015.

[10] S. H. Jung, J. C. Kim, and C. B. Sim, "Prediction Data Processing Scheme using an Artificial Neural Network and Data Clustering for Big Data," *Inter. J. of Ele.Com. Eng.*, Vol.6, No.1, pp.330-336, 2016.



**정 세 훈**

e-mail : iam1710@hanmail.net
2010년 순천대학교 멀티미디어공학과(학사)
2012년 순천대학교 멀티미디어공학과(석사)
2017년 순천대학교 멀티미디어공학과(박사)
2015년~현 재 광양만권 SW융합연구소 팀장
관심분야 : 빅데이터 처리 및 확률 분석, 데이터마이닝



**신 창 선**

e-mail : csshin@sunchon.ac.kr
1996년 우석대학교 전산학과(학사)
1999년 한양대학교 컴퓨터교육과(석사)
2004년 원광대학교 컴퓨터공학과(박사)
2005년~현 재 순천대학교 정보통신·멀티미디어공학부 교수
관심분야 : 분산컴퓨팅, 실시간 객체모델, 기계학습, 시계열분석



**조 용 윤**

e-mail : yycho@sunchon.ac.kr
1995년 인천대학교 전산학과(학사)
1998년 숭실대학교 컴퓨터학과(석사)
2006년 숭실대학교 컴퓨터학과(박사)
2009년~현 재 순천대학교 정보통신·멀티미디어공학부 교수
관심분야 : 시스템 소프트웨어, 유비쿼터스 컴퓨팅, 기계학습



**박 장 우**

e-mail : jwpark@sunchon.ac.kr
1989년 한양대학교 전자공학과(학사)
1991년 한양대학교 전자공학과(석사)
1993년 한양대학교 전자공학과(박사)
1995년~현 재 순천대학교 정보통신·멀티미디어공학부 교수
관심분야 : SoC, USN, 기계학습, 시계열분석

박 명 혜

e-mail : myunghye.park@kepco.co.kr
1993년 경북대학교 전자공학(학사)
1995년 경북대학교 전자공학(석사)
1995년~현 재 한전 전력연구원
책임연구원
관심분야 : 유·무선 통신망 설계, IoT

김 영 현

e-mail : younghyun.kim@kepco.co.kr
2002년 한국항공대학교 통신정보공학(학사)
2004년 광주과학기술원 정보통신공학(석사)
2004년~현 재 한전 전력연구원
선임연구원
관심분야 : 유·무선 통신망 설계, 사물인터넷

이 승 배

e-mail : sblee83@kepco.co.kr
1993년 청주대학교 행정학(학사)
1996년 충북대학교 행정학(석사)
1993년~현 재 한전 전력연구원
책임연구원
관심분야 : 전력자동화통신망, 사물인터넷

심 춘 보

e-mail : cbsim@sunchon.ac.kr
1996년 전북대학교 컴퓨터공학과(학사)
1998년 전북대학교 컴퓨터공학과(석사)
2003년 전북대학교 컴퓨터공학과(박사)
2005년~현 재 순천대학교 정보통신·
멀티미디어공학부 교수
관심분야 : 멀티미디어 DB, 데이터 분석, 빅데이터 처리 및 분석