지진 분석을 위한 결합 손실 함수를 활용한 심층 군집화 모델 개선 연구

김병학*, 안재광, 황의홍
*기상청 지진화산연구과

*bhkim715@korea.kr, propjk@korea.kr, hkawon1@korea.kr

A Study on the Improvement of Deep Clustering Model using Cohesive Loss Function for Earthquake Analysis

Kim Byeong-Hak*, Ahn Jae-Kwang, Hwang Eui-Hong
*Earthquake and Volcano Research Division, Korea Meteorological Administration

요 약

본 연구는 지진계의 기록을 이미지화 하여 비지도학습 기반의 신호 분류를 수행하고자 심층 분리형 군집화 모델을 개발하였다. 이때 사용되는 심층 분리형 군집화 모델은 기존 군집화 알고리즘에 딥러닝 모델이 추가되었으며 네 가지의 모델 손실함수를 결합한 전역 손실함수로 모델을 최적화한다. 이로 인해 어느 한 쪽의 손실에 치우치지 않은 심층학습 모델로 레이블이 없는 이미지 분류가 가능해진다. 따라서, 새로운 심층 분리형 군집화 방식은 동일 조건의 실험에서 기존 DEC[1] 모델보다 ACC 평가 기준 15.8%, NMI 평가 기준 19.9% 성능을 능가했다.

Ⅰ. 서 론

심층 군집화(Deep Clustering)는 기존의 군집화 알고리즘에 딥러닝 모델을 적용한 방식이다. 주로 이 방법은 raw data 에 대해 군집화하는 대신 딥러닝 모델로 미리 입력 데이터의 특징을 학습하는 전처리 과정을 포함하고 있어 다차원을 가진 데이터의 군집화에 많이 사용된다. 그러나 분포된 데이터의 전체적인 분산이 크거나 군집 간 중첩현상(Overlapping)에 있어서는 데이터를 잘 분류해내지 못하는 한계점을 가지고 있다. 따라서, 본 연구에서는 지진파 이미지를 분류하기 위해 4 가지의 이미지 데이터 집합(MNIST, MNIST-test, Fashion MNIST, USPS)을 대상으로 한 심층 군집화 알고리즘을 제안하고자 한다.

Ⅱ. 본론

본 연구에서는 글로벌 최적화를 위해 4 개의 손실 함수를 사용하여 군집 간의 분리를 기반으로 하는 혁신적인 심층 군집화 알고리즘을 제안한다. 본 장에서는 적용된 4 가지의 손실 함수에 대하여 기술한다.

1. 합성곱 오토 인코더를 사용한 특징 추출

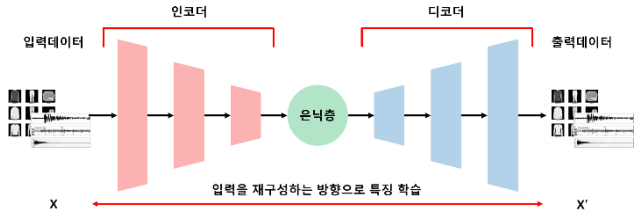


그림 1 합성곱 오토인코더에 의한 특징 학습 구조도

첫 번째 손실함수 L_r은 오토인코더의 재구성 함수이다. **그림 1** 에 합성곱 오토 인코더를 활용한 특징학습 구조를 나타내었다. 손실함수 L_r은 다음과 같다.

$$L_r = \frac{1}{m}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2^2 \quad (1)$$

여기서 m은 전체 데이터 샘플 개수, i는 샘플 인덱스, $\boldsymbol{x}$는 입력 데이터 자체를 의미한다.

2. 데이터의 군집 간 확률분포를 평가

두 번째 손실함수 L_c 는 **Ⅱ-1** 에서 학습된 특징을 사용해 k-평균 군집화[2]를 수행하고 데이터 포인트의 군집 간 확률분포 차이를 손실함수로 계산한다. 해당 손실 함수는 Student's t-분포[3]에 의해 계산되며 다음과 같이 구성된다.

$$L_c = D_{KL}(\boldsymbol{P}||\boldsymbol{Q}) = \sum_i \sum_j p_{ij} \log p_{ij}/q_{ij} \quad (2)$$

여기서 D_{KL}은 Kullback-Leibler Divergence (KLD)[4]를 의미하고 p_{ij}와 q_{ij}는 각각 다음과 같다.

$$q_{ij} = \frac{\left(1 + \|\boldsymbol{z}_i - \boldsymbol{\mu}_j\|^2\right)^{-1}}{\sum_{j'}\left(1 + \|\boldsymbol{z}_i - \boldsymbol{\mu}_j\|^2\right)^{-1}} \quad (3)$$

$$p_{ij} = \frac{Q^T/\sum_i q_{ij}}{\sum_{j'}\{Q^T/\sum_i q_{ij'}\}} \quad (4)$$

여기서 j는 샘플이 속하게 될 군집(cluster)을 뜻하고, $\boldsymbol{z}$는 합성곱 오토인코더로 학습된 은닉층의 특징, $\boldsymbol{\mu}$는 군집의 중심점 즉 centroid 를 의미하며, $Q = (e^{q_{ij}} - 1)/$

$(e-1)$, $T>0$이다. 본 연구에서 우리는 T의 값을 3 으로 설정하였고 확률값에 맞도록 Q^T 의 범위는 0 에서 1 사이가 된다.

q_{ij} 는 i 번째 샘플이 j 번째 클러스터에 속할 확률을 나타내는 soft 분포이며 Q^T 의 비선형성을 강조하여 만들어진 p_{ij} 를 타겟 분포로 설정하여 반복적으로 손실값이 계산된다. p_{ij} 는 군집들 간 손실함수 기여도의 불균형을 완화시킨다.

3. 중앙 손실함수를 활용한 군집 내 응집력 향상

군집화 알고리즘의 최후의 목표는 각 모델이 정의한 유사도의 기준에 따라서 상대적 유사도가 높은 샘플은 같은 군집에 속하게 하고 유사도가 낮은 샘플은 다른 군집에 속하게 하는 것이다. 본 연구에서는 중앙 손실함수[5]를 활용하여 동일 군집에 속한 유사한 샘플들 간의 분산을 감소시켰다. 중앙 손실함수는 다음과 같다.

$$L_W = \frac{1}{2}\sum_{i=1}^{m}\left\|\boldsymbol{z}_i - \boldsymbol{\mu}_{y_i}\right\|_2^2 \tag{5}$$

여기서 y_i 는 모델이 예측한 i 번째 샘플에 대한 군집 라벨을 의미하고 $\boldsymbol{\mu}_{y_i}$ 는 y_i 번째 군집의 centroid 이다. $\boldsymbol{\mu}_{y_i}$ 는 반복적인 학습과정을 통해 지속적으로 업데이트된다.

4. 제안하는 손실함수를 활용한 군집 간 분리도 향상

본 세션에서 소개하는 L_B 는 서로 다른 군집 간의 코사인 유사거리 (inter-cluster cosine distance)를 변형하여 계산된 새로운 손실함수이다.

$$L_B = \frac{1}{2}\cdot {}_nC_2 \cdot \sum_{j=0}^{n-1}\sum_{k=0}^{n-1}\left[\log\left\{ReLU\left(\frac{\boldsymbol{\mu}_j\cdot\boldsymbol{\mu}_k}{\|\boldsymbol{\mu}_j\|_2\|\boldsymbol{\mu}_k\|_2}\right)+1+\epsilon\right\}\right] \tag{6}$$

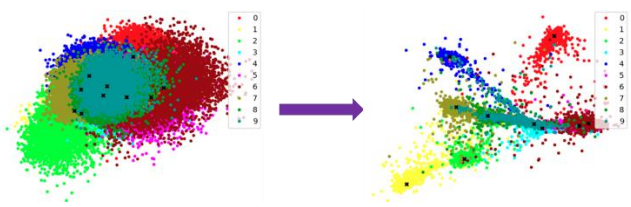
n 은 클러스터의 총 개수를 의미하고 ${}_nC_2$ 는 n 개 클러스터들을 2 개씩 짝을 지을 때의 경우의 수를 구하기 위하여 계산한다. 여기서 활성화 함수 ReLU 는 군집 쌍 조합의 코사인 유사도 거리 $\left(\boldsymbol{\mu}_j\cdot\boldsymbol{\mu}_k / \|\boldsymbol{\mu}_j\|_2\|\boldsymbol{\mu}_k\|_2\right)$ 값이 음의 범위를 갖는 것을 제한하는 역할을 한다. 비슷한 목적으로 1 과 ϵ은 각각 로그 함수의 최종 출력이 음수와 0 의 값을 갖는 것을 방지하기 위해 사용된다.

마지막으로 우리는 세션 Ⅱ에서 소개한 4 가지의 손실함수를 하나의 전역함수로 가중 파라미터 곱을 하여 모델을 최적화하는 데 사용했다[6].

$$L = \alpha\cdot L_r + \beta\cdot L_c + \gamma\cdot L_W + \omega\cdot L_B \tag{7}$$

$\alpha,\beta,\gamma,\omega$ 각각의 값은 모델의 하이퍼 파라미터로서 경험적인(heuristic) 방식으로 결정할 수 있다.

Ⅲ. 실험결과



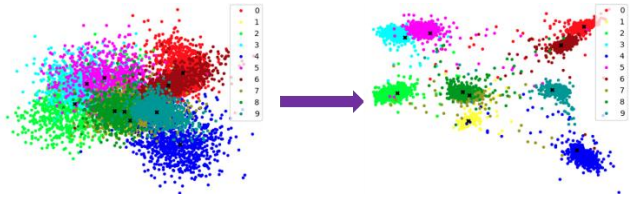


그림 2 제안하는 군집 알고리즘을 적용한 전후 결과. (위) MNIST 데이터 세트, (아래) USPS 데이터 세트

본 논문에서 우리는 심층 군집화를 위한 새로운 접근법을 제안했다. **그림 2** 로 시각화된 실험결과는 우리가 제안한 알고리즘이 레이블이 없는 데이터 세트를 효과적으로 클러스터링할 수 있다는 것을 보여주었다. 제안된 군집화 방식은 재구성, 클러스터 내 동질성, 클러스터 간 분리성 및 KL-발산 손실 함수를 포함하여 기존 군집화 알고리즘의 성능을 Clustering accuracy (ACC) 기준 15.8%, Normalized Mutual Information (NMI) 기준 19.9% 이상 능가하였다. 또한, 제안한 방식은 손실 가중치를 조정하여 태스크에 따라 높은 성능으로 모델을 최적화할 수 있다는 점에서 효과적이다.

Ⅳ. 결론

본 연구에서 제안한 심층 분리형 군집화 모델은 클러스터링 결과 중첩된 군집 분포에 대한 개선효과를 볼 수 있었다. 이는 L_W와 L_B에 대한 영향으로 판단된다. 따라서, 본 연구에서 개발된 알고리즘을 통해 지진 식별 성공성에 대한 평가를 수행하고 보완해 가고자 한다.

ACKNOWLEDGMENT

This study was supported by the “Development of earthquake, tsunami, volcano monitoring and prediction technology (KMA2018-00820)” project of the Korea Meteorological Administration.

참 고 문 헌

[1] Xie, J., Girshick, R., & Farhadi, A. “Unsupervised deep embedding for clustering analysis.” In International conference on machine learning, p. 478-487, 2016.

[2] J. Macqueen, et al., “Some methods for classification and analysis of multivariate observations”, Proc. fifth Berkeley symposium on mathematical statistics and probability, pp.281-297, 1967.

[3] Student, “The probable error of a mean”, Biometrika, vol.6 no.1, pp.1-25, 1908.

[4] S. Kullback and R.A. Leibler, “On information and sufficiency”, The annals of mathematical statistics, vol.22, no.1, pp.79-86, 1951.

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition”, European conference on computer vision, Springer, Cham, pp.499-515, 2016.

[6] KIM, Byeonghak, et al. “Deep Clustering for Improved Inter-Cluster Separability and Intra-Cluster Homogeneity with Cohesive Loss”, IEICE TRANSACTIONS on Information and Systems, Vol. E104-D no.5, pp.776-780, 2021.