Journal of Korea Multimedia Society Vol. 23, No. 1, January 2020(pp. 1-7)
https://doi.org/10.9717/kmms.2020.23.1.001

합성곱 오토인코더 기반의 응집형 계층적 군집 분석

박노진†, 고한석††

Agglomerative Hierarchical Clustering Analysis with Deep Convolutional Autoencoders

Nojin Park†, Hanseok Ko††

ABSTRACT

Clustering methods essentially take a two-step approach; extracting feature vectors for dimensionality reduction and then employing clustering algorithm on the extracted feature vectors. However, for clustering images, the traditional clustering methods such as stacked auto-encoder based k-means are not effective since they tend to ignore the local information. In this paper, we propose a method first to effectively reduce data dimensionality using convolutional auto-encoder to capture and reflect the local information and then to accurately cluster similar data samples by using a hierarchical clustering approach. The experimental results confirm that the clustering results are improved by using the proposed model in terms of clustering accuracy and normalized mutual information.

Key words: Clustering Analysis, K-means Clustering, Convolutional Autoencoder, Hierarchical Clustering

1. 서 론

최근 유행한 아프리카 돼지 열병바이러스(ASFV), 신종 인플루엔자(H1N1) 등 여러 감염병으로 인해 각 국가는 막대한 경제적 사회적 손실을 보았다. 이러한 피해를 예방하기 위해 각 나라는 다양한 노력을 기울여 왔고 국내 또한 질병관리본부를 설립하여 여러 연구를 시행해 피해 대책을 해왔지만, 감염병 피해는 지속적으로 발생해 왔다. 예컨대 감염병 피해를 막기 위한 노력으로는, 빅데이터와 딥러닝 기술을 활용한 예측모델[1], 도메인 지식 기반 해외 발생 감염병 국내 유입가능성 탐지 모델[2]가 대표적이다. 하지만, 획득 가능한 데이터는 급격히 증가하였음에도 불구하고, 실제 방역 문제를 해결하기 위한 적절한 형태의 정보를 제한된 시간 이내에 얻는 것이 어렵기 때문에 사전적으로 유사 데이터끼리 군집화(clustering)하는 것이 중요하다. 이때 군집화하는 데이터가 이미지와 같은 고차원 데이터일 경우 전통적인 군집화 방식만을 사용하기에는 효율적이지 않으며 공간 정보 또한 무시해서는 안된다. 하지만 기존 방법[3,4,5,6,7,8]은 완전 연결 계층(fully connected layer)로 구성된 적층 오토인코더(stacked auto encoder)방식을 사용하여 고차원의 데이터를 저차원으로 치환시킬 때, 공간 정보를 반영하지 않았다. 이에 본 논문에

※ Corresponding Author : Han Seok Ko, Address: 145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea, TEL : +82-2-3290-3239, FAX : +82-2-3291-2450, E-mail : hsko@korea.ac.kr
Receipt date : Nov. 12, 2019, Revision date : Dec. 5, 2019
Approval date : Dec. 16, 2019

† School of Electrical Engineering, Korea University (E-mail : njpark@ispl.korea.ac.kr)
†† School of Electrical Engineering, Korea University
※ This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grantnumber : H18C1234).

서는 고차원 감염병 데이터 간 유사성을 찾아내기 위해 첫 번째로, 합성곱-오토인코더를 활용함으로써 데이터 압축을 시행할 때 공간 정보를 반영할 수 있도록 한다. 이후 압축된 데이터가 존재하는 코딩층(code-layer)에서 응집형 계층적 병합군집(agglomerative hierarchical clustering)을 수행하여 유사한 데이터끼리 군집화를 한다. 제안한 방법의 객관적 평가를 위해 최근 관련된 연구들과 결과를 비교함으로써, 압축된 정보를 활용하여 군집화한 결과가 기존 방법[3,4]등에 비하여 성능이 우수함을 증명한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 문제 제기를 통해 기존의 k-means와 stacked autoencoder의 방법으로는 해결하지 못한 공간적 정보 손실 해결과 효과적 군집화의 필요성을 정리하고, 3장에서 제안한 모델을 설명한다. 그리고 4장에서는 제안한 모델로 실험한 결과를 토대로 기존 모델들과의 성능 비교를 하고 5장에서 결론을 맺는다.

2. 문제 제기

2.1 인공신경망 구조적 공간정보 손실

비지도 학습(unsupervissd learning)을 위한 인공신경망 구조 중 하나인 합성곱-오토인코더는 적층 오토인코더가 공간 정보를 무시하는 경향을 보완한 방법이다. 완전 연결 계층으로 이루어진 적층 오토인코더는 입력 데이터를 1차원으로 치환하여야 하는 문제점이 있기 때문에, 이미지와 같이 공간 정보를 가지는 경우 공간 정보가 소실되며 서로 다른 각 층(layer)의 모든 노드(node)가 연결되어 있으므로 공간적 구조(spatial structure)를 보존하지 못하는 큰 단점이 존재한다. 이에 반해 합성곱-오토인코더는 입/출력 모두 2차, 3차 혹은 4차원 데이터로 처리하기 때문에 공간 정보를 유지할 수 있는 큰 장점이 있다(Fig. 1). 합성곱 오토인코더는 적층 오토인코더와 같이 데이터의 차원을 줄이는 인코더(recognition network)와 입력 데이터를 복원시키는 디코더(generative network)로 구성되어있다. 구체적으로 합성곱-오토인코더의 출력층의 차원은 입력층의 차원과 동일한 구조이며, 은닉층의 차원은 입력층의 차원보다 작다. 따라서, 은닉층은 입력 데이터의 불필요한 특징을 제거한 압축된 특징을 학습할 수 있다. 일반적인 합성곱-오토인코더의 인코더는 입력 데이터 x를 점차 더 작은 차원인 은닉층 h로 매핑(mapping)시키는 역할을 하며, 이는 식(1)과 같다.

$$h = f_W(x) = \sigma(W * x) \tag{1}$$

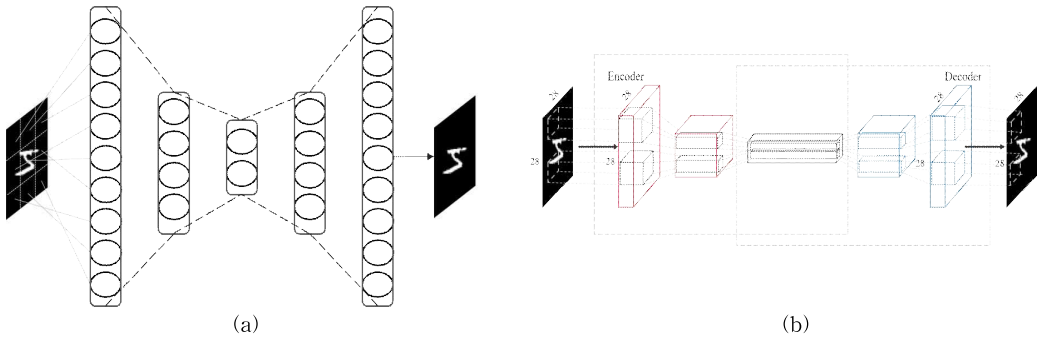
위의 식에서 $f(\bullet)$는 인코더를 나타내며, x는 입력벡터, W는 인코더의 파라미터, σ는 ReLU등과 같은 활성화 함수(activation function), $*$는 합성곱 연산자(convolution operator)이다.

디코더는 인코더로부터 압축된 데이터가 있는 은닉층 h로부터 입력을 받아 x와 유사할 수 있도록 재구성하는 역할을 수행하며 이는 식(2)와 같다.

$$x' = g_U(h) = \sigma(U * h) \tag{2}$$

위의 식에서 $g(\bullet)$는 디코더를 나타내며, U는 디코더의 파라미터, x'은 디코더의 출력값이다.

최종적으로 합성곱-오토인코더는 모델의 출력이 입력 데이터에 최대한 근사하는 방향으로 재구성 손실(reconstruction error)을 최소화시킴으로써 입력



(a) (b)

Fig. 1. Comparison between (a) stacked auto-encoder and (b) convolutional auto-encoder on the error propagation capturing local information. Stacked auto-encoder's error is propagated to all parameters while convolutional auto-encoder propagates error only to its masked region.

데이터의 핵심적인 특징(feature)을 학습하도록 한다. 학습을 위한 목적 함수는 mean squared error (MSE)를 사용하여 아래와 같이 정의된다.

$$L_{CAE} = \frac{1}{n}\sum_{i=1}^{n} \| g_U(f_W(x_i) - x_i) \|_2^2 \quad (3)$$

위의 식에서 n은 데이터 수이고 x_i는 i번째 데이터이다.

2.2 군집화의 효율성

비지도학습인 군집화(clustering)는 각 개체의 그룹정보(label) 없이 유사성을 측정하여 유사한 데이터끼리 그룹화 하는 것이 목적이며, 유형으로는 크게 계층적 군집화(hierarchical clustering)와 분할적 군집화(partitional clustering)알고리즘이 있다(Fig. 2).

분할적 군집화 알고리즘은 데이터 세트를 k 개의 그룹 수로 세분화하는 클러스터링 기법이다. 이 방법으로는 각 군집의 중심으로부터 가까운 데이터를 묶는 *k-means* clustering(MacQueen 1967)대표적이다.

$$\min_{b,w} J = \sum_{n=1}^{N}\sum_{k=1}^{K} w_{nk} \| x_n - \mu_k \|_2^2 \quad \text{s.t.} \sum_k w_{nk} = 1, \forall k \quad (4)$$

여기서 N은 데이터 수, K는 군집 수, μ_k는 군집 K에 따른 중심, w_{nk}는 n번째 데이터가 k번째 군집에 속하는지를 나타내는 이진 변수이며, 모든 데이터는 하나의 군집에 속해야 한다.

이에 반해, 계층적 군집분석은 군집 수를 지정하지 않아도 각 개체들이 결합 또는 분할되는 과정을 나타내는 트리 형태의 구조인 덴드로그램(dendrogram)을 통해 적절한 군집수를 나눌 수 있는 장점이 있다. 구체적으로 계층적 군집분석은 개체 간의 유사성을 이용하여 가장 유사한 데이터끼리 합병해 나가는 응집형 계층적 군집(agglomerative hierarchical clustering)과 유사성이 없는 데이터끼리 점차 분할해 나가는 분할형 계층적 군집(divisive hierarchical clustering)이 있다. 군집간의 거리는 대표적으로 단일연결법, 완전연결법, 메디안(median)연결법, 와드(ward)연결법 등이 있다. 단일연결법은 군집간의 최단거리를 이용함으로써 상이한 군집을 찾는데 중점을 둔 방법이며, 완전 연결법은 군집간 최장거리를 이용함으로써 응집성에 중점을 둔 군집법이다. 또한 메디안 연결법은 중심연결법에서 작은 군집이 무시되는 경향을 보완한 방법이며, 와드 연결법은 군집간 거리와 군집 내 편차 제곱합에 근거하여 병합해 나가는 방식으로써 비슷한 크기의 군집끼리 병합하는 특징이 있다. 따라서 계층적 군집분석은 거리의 정의에 따라 다른 군집으로 형성될 수 있으므로 자신의 목적에 부합한 거리를 정의해야만 한다. 합성곱-오토인코더와 군집 분석의 이론적 배경을 토대로 본 논문의 3장에서 효율적인 군집화 방법을 소개한다.

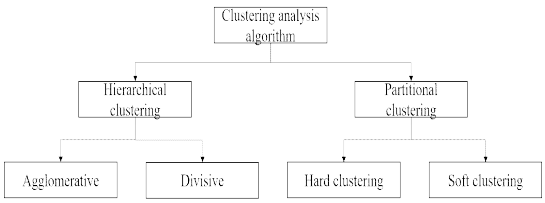


Fig. 2. The set of hierarchical clusters is nested with overlapping subsets that are organized as a tree. The partitional clustering is a division of the set of data samples into non-overlapping subsets such that each data sample is in exactly one subset.

3. 제안한 방법

본 논문에서는 고차원의 데이터를 저차원으로 효율적으로 줄이며 서로 유사한 데이터끼리 군집화하는 방법을 제안한다. 제안한 방법은 CNN (convolutional neural networks)의 유형 중 하나인 합성곱-오토인코더의 인코더 부분을 보다 깊게 convolution과 sub-sampling을 수행하여 강력한 특징을 학습할 수 있도록 하며, 압축된 정보가 존재하는 코딩층(code-layer)에서 와드 연결법(Ward linkage)을 통한 응집형 계층적 군집을 수행함으로써 노이즈(noise)와 이상치(outlier)에 강인할 수 있도록 하였다. 본 논문에서 제안한 자세한 모델은 Fig. 3에 자세히 시각화하였다.

전통적인 군집화 방식만을 사용하기에는 이미지 데이터는 고차원이기 때문에 완전 연결 계층으로 이루어진 적층 오토인코더를 이용하여 저차원으로 치환시킨 후 군집화를 시행하는 기존 연구[3,4,5,6,7,8]가 있지만 공간 정보를 반영하지 않았는다는 점과 *k- means* clustering은 이상치(outlier)에 민감한 단점이 있다. 본 논문에서는 이러한 단점을 개선하기 위해 합성곱-오토인코더 개념 및 와드 연결법을 통한 군집화를 하였다. 합성곱-오토인코더는 기존 적

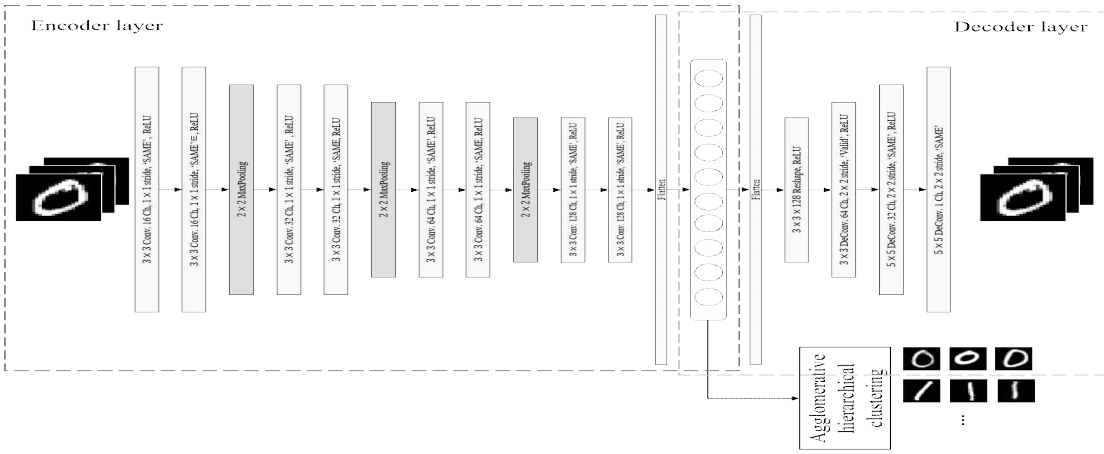


Fig. 3. Proposed encoder reinforced CAE based agglomerative hierarchical model for clustering.

층 오토인코더에 합성곱신경망(covolutional nearal networks)의 개념을 추가한 것으로 출력 데이터를 입력 데이터와 가능한 동일하게 재구성(reconstruction)하는 방향으로 학습이 진행된다. 이 과정에서 코딩층의 특징 데이터는 합성곱-오토인코더의 입력 데이터를 디코더가 재구성할 수 있게 핵심적인 정보들을 압축되어 만들어진다. 제안된 모델은 기존에 합성곱 오토인코더 개념을 활용한 연구[9]보다 인코더(recognition network)층을 더욱 깊게 쌓아 올려서 코딩층이 더욱 강인한 특징을 학습할 수 있도록 모델을 구축하는 방식을 사용하였다. 즉, 기존 합성곱-오토인코더의 인코더 부분을 여러 개의 convolution과 sub-sampling을 반복적으로 수행함으로써 불필요한 파라미터 수는 최대한 줄이고 보다 핵심적인 정보를 강화하여 최종적으로 코딩층에서 보다 강인한 특징을 학습할 수 있도록 모델을 구축하였다.

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2} \tag{5}$$

또한, 합성곱-오토인코더의 코딩층에서 응집형 계층적 군집방식(agglomerative hierarchical clustering) 사용하기 위해 유클리드 거리(euclidean distance)를 이용하여 각 개체 간의 거리를 측정하였고, 응집형 계층적 군집을 형성할 때 비교적 정확한 결과가 나오는 와드 연결법(Ward linkage)방식을 사용하여 가장 유사한 군집들을 병합할 수 있도록 하였다. 와드 연결법은 군집 간 거리만으로 데이터를 연결하는 것이 아닌, 군집 내 편차들의 제곱합(error sum of square)에 근거하여 병합해 나가는 방법으로써 노이즈(noise)나 이상치(outlier)에 덜 민감한 장점이 있다.

$$Ward\,Distance = \sum_{i \in A \cup B} \| x_i - m_{A \cup B} \|^2 - \left\{ \sum_{i \in A} \| x_i - m_A \|^2 + \sum_{i \in B} \| x_i - m_B \|^2 \right\} \tag{6}$$

위 식에서 m은 군집의 평균, A, B 는 군집을 나타낸다. 즉, 위의 식은 두 군집이 하나의 군집으로 형성되었을 때의 평균과 각 개체들과의 거리, 또 서로 다른 두 군집의 평균과 그에 속하는 개체들과의 거리의 차이이다. 병합된 군집의 오차 제곱 합은 병합 이전의 합 보다 커지게 되는데, 이 증가량이 작아지는 방식으로 군집화 시행하게 되기 때문에 비슷한 크기의 군집이 형성되며 계층적 군집분석에서 비교적 정확한 결과를 도출할 수 있다.

4. 실험 결과

제안한 방식과 기존의 방식간의 비교분석을 위해 MNIST dataset을 사용하였다. MNIST 데이터는 60,000장의 트레이닝 데이터와 10,000장의 테스트 데이터로 이루어진 데이터셋으로 0~9 사이의 28×28 필기체 이미지로 구성되어있다. 본 논문에서 제안한 합성곱-오토인코더의 인코더(recognition network)층을 여러개의 convolution과 sub-sampling을 반복적으로 수행함으로써 코딩층에서 강인한 특징을 학습할 수 있도록 모델을 구축하고 계층적 군집분석을 한 결과 기존의 방식들보다 성능이 우수하였음을 확인하였다. Fig. 4는 제안한 합성곱-오토인코더와 계층적 군집화의 성능을 시각화하기 위해 t-SNE[10]

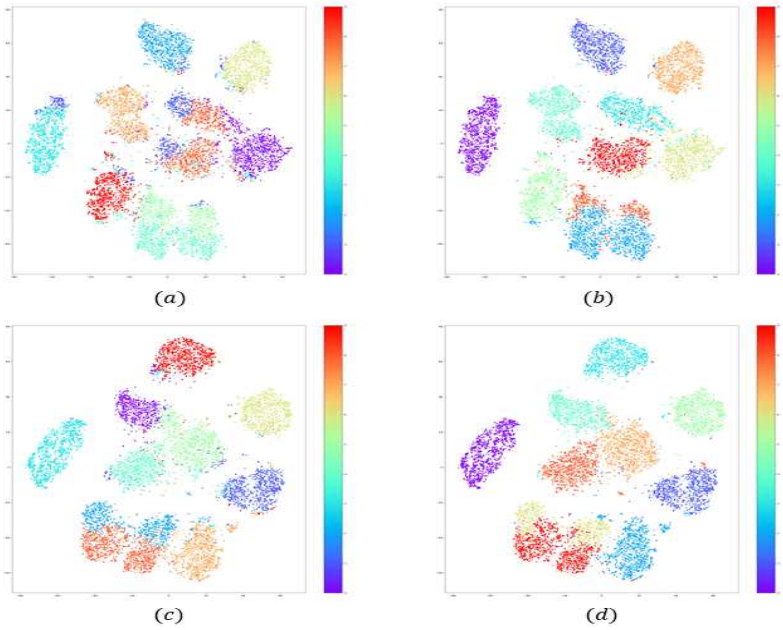


Fig. 4. Visualization of clustering results of MNIST-TEST. Different colors mark different clusters. (a) is CAE+ *k-means* clustering, (b) is CAE+Hierarchical clustering, (c) is proposed the CAE+ *k-means* clustering and (d) is the proposed method. It can be seen that the proposed method shows the best performance.

를 사용하여 나타낸 것이다. t-SNE는 고차원의 데이터를 저 차원의 데이터로 거리 관계를 유지하며 임베딩(embedding) 시키는 알고리즘이다.

군집성능평가를 하기 위해 성능평가지표로는 clustering accuracy(ACC)와 normalized mutual information(NMI)를 사용하였다. ACC와 NMI는 다음과 같다.

$$ACC = \sum_{i=1}^{N} \frac{\delta(r_i, map(c_i))}{N} \tag{7}$$

위에서의 N은 데이터 수, c_i와 r_i는 예측된 군집 레이블 및 실제 정답 군집, $map(c_i)$는 각 예측된 군집 레이블을 실제 레이블과 동일하게 매핑하는 함수이다. 마지막으로 $\delta(r_i, map(c_i))$는 $r_i = map(c_i)$가 성립할 시 1이고 성립하지 않을시 0인 델타함수이다.

$$NMI(Y, C) = \frac{MI(Y, C)}{mean[H(Y) + H(C)]} \tag{8}$$

위에서 Y는 실제 레이블, C는 군집 레이블, H는 엔트로피이며, MI는 mutual information metric이다. Normalized mutual information은 mutual information 값이 0과 1의 사이 값이 되도록 정규화한 지표이며, 이 때 upper bound는 Y와 C가 가진 엔트로피(불확실성)의 산술평균값 혹은 기하평균, 최대/최소값 등을 사용할 수 있다.

제안된 합성곱-오토인코더(CAE+hierarchical: 95.63%)의 강인한 특징 추출로 인해, 코딩층에서의 핵심적인 특징을 가지고 계층적 군집분석을 한 결과 Table 1에 나타낸 바와 같이 기존의 방법 (CAE+k-means: 84.90%)들에 비해 약 10% 향상을 가진 가장 우수한 성능을 보였다.

5. 결 론

본 논문에서는 합성곱-오토인코더의 인코더를 강화하여 코딩층에서 강인한 특징을 학습할 수 있도록 모델을 구축하고 계층적 군집화를 수행하여 유사한 데이터끼리 군집화하는 방법을 제안하였다. 실험결과 기존의 방식보다 우수한 성능향상이 있었으며 잉여정보를 제거하고 중요한 정보만을 사용하는 방식이 군집화 성능향상에 기인한다는 것을 확인하였다. 이 과정에 있어 각 대응되는 인코더와 디코더에 skip-connection을 활용하는 연구도 진행해 보았지만, 오히려 데이터 압축에 악영향을 끼침을 확인하였다. 다시말해, 인코딩 과정에서의 정보 손실을 고려하여 디

Table 1. Comparison of clustering performance. MNIST-full consists of total 70,000 handwritten digits of 28x28 pixels. MNIST-test containing 10,000 images(MNIST-test set)

Methods	Dataset			
	MNIST-full		MNIST-test	
	ACC [%]	NMI [%]	ACC [%]	NMI [%]
k-means	54.24	48.52	54.63	50.18
SEC[5]	80.37	-	-	-
SAE+*k*-means	78.17	71.46	66.81	59.59
DEC[4]	84.08	81.28	69.94	67.69
IDEC	84.21	83.81	71.45	69.40
CAE+*k*-means	84.90	79.27	79.00	72.55
DEC-conv	88.63	87.59	84.83	82.62
DCEC[9]	88.97	88.49	85.29	83.61
proposed CAE+*k*-means	87.82	85.12	81.62	76.59
CAE+hierarchical	92.38	88.48	81.05	79.52
proposed CAE + hierarchical	95.63	91.92	87.53	84.10

코더에 직접적으로 연결을 할 때, 재구성 손실(reconstruction error)은 확실히 줄지만 오히려 오토인코더의 가장 중요한 목표인 코딩층에 중요한 정보가 남지 않게 되어 본 논문의 군집화 목적에는 적합하지 않았다. 향후 연구로는 계층적 군집 분석의 시간복잡도 및 메모리 요구량의 단점을 극복하기 위해 새로운 군집분석 방식을 적용해볼 예정이다.

REFERENCES

[1] S.H. Kim, J.K. Choi, J.S. Kim, A.R. Jang, J.H. Lee, K.J. Cha, et al., "Animal Infectious Diseases Prevention through Big Data and Deep Learning," *Journal of Intelligence and Information System*, Vol. 24, No. 4, pp. 137-154, 2018.

[2] M.N. Hwang and S.W. Lee, "Socio-national Issues Detection Modeling based on Domain Knowledge-focusing on the Issue of Increase in Domestic Inflow Infectious Diseases," *The Journal of the Korea Contents Association,* Vol. 17, No. 12, pp. 158-168, 2017.

[3] C. Song, F. Liu, Y. Huang, L. Wang, and T. Tan, "Auto-encoder Based Data Clustering," *Proceeding of Iberoamerican Congress on Pattern Recognition,* pp. 117-124, 2013.

[4] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," *Proceeding of International Conference on Machine Learning*, pp. 478-487, 2016.

[5] F. Nie, Z. Zeng, I.W. Tsang, D. Xu, and C. Zhang, "Spectral Embedded Clustering: A Framework for In-sample and Out-of-sample Spectral Clustering," *IEEE Transactions on Neural Networks*, Vol. 22, No. 11, pp. 1796-1808, 2011

[6] X. Peng, J. Feng, J. Lu, W.Y. Yau, and Z. Yi, "Cascade Subspace Clustering," *Proceeding of Thirty-First Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 2478-2484, 2017.

[7] X. Peng, S. Xiao, J. Feng, W.Y. Yau, and Z. Yi, "Deep Subspace Clustering with Sparsity Prior," *Proceeding of International Joint Conference on Artificial Intelligence*, pp. 1925-1931, 2016.

[8] H.J. Lee, "Performance Improvement of Deep Clustering Networks for Multi Dimensional Data," Journal of Korea Multimedia Society, Vol. 21, No. 8, pp. 952-959, 2018

[9] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep Clustering with Convolutional Autoencoders," *Proceeding of International Conference on*

Neural Information Processing, pp. 373-382, 2017.

[10] L.V.D. Maaten and G. Hinton, "Visualizing Data Using Accelerating t-SNE Using Tree-Based Algorithms," *Journal of Machine Learning Research*, Vol. 9, pp. 2579-2605, 2008.

박 노 진

2019년 백석대학교 소프트웨어학 공학사 취득
2019년 현재 고려대학교 전기전자공학과 석사과정
관심분야 : 영상신호처리, 패턴인식, 머신러닝

고 한 석

1982년 Carnegie-Mellon Univ. 전기공학 공학사 취득
1988년 Johns Hopkins Univ. 전자공학 공학석사 취득
1992년 Catholic Univ. of America 전자공학 공학박사 취득

2015년 현재 고려대학교 전기전자공학과 교수
관심분야 : 영상/음성 신호처리, 패턴인식