

# 비지도학습 데이터의 정확성 측정을 위한 클러스터별 분류 평가 예측 모델에 대한 연구

정세훈<sup>†</sup>, 김종찬<sup>\*\*</sup>, 김치용<sup>\*\*\*</sup>, 유강수<sup>\*\*\*\*</sup>, 심춘보<sup>\*\*\*\*\*</sup>

## A Study on Classification Evaluation Prediction Model by Cluster for Accuracy Measurement of Unsupervised Learning Data

Se Hoon Jung<sup>†</sup>, Jong Chan Kim<sup>\*\*</sup>, Kim Cheeyong<sup>\*\*\*</sup>,  
Kang Soo You<sup>\*\*\*\*</sup>, Chun Bo Sim<sup>\*\*\*\*\*</sup>

### ABSTRACT

In this paper, we are applied a nerve network to allow for the reflection of data learning methods in their overall forms by using cluster data rather than data learning by the stages and then selected a nerve network model and analyzed its variables through learning by the cluster. The CkLR algorithm was proposed to analyze the reaction variables of clustering outcomes through an approach to the initialization of K-means clustering and build a model to assess the prediction rate of clustering and the accuracy rate of prediction in case of new data inputs. The performance evaluation results show that the accuracy rate of test data by the class was over 92%, which was the mean accuracy rate of the entire test data, thus confirming the advantages of a specialized structure found in the proposed learning nerve network by the class.

**Key words:** Cluster Verification, RNN, LSTM, Classification, Unsupervised Learning, Clustering  
*k*-LSTM RNN

### 1. 서 론

새롭게 생성되는 대용량 다차원 정보의 저장 및 검색 관련 분야에서 다차원 정보에 내재되어 있는 유용한 지식을 검색(Retrieval)하고 새로운 지식을 발견하기 위해 빅데이터 마이닝이라는 분야가 각광 받고 있다. 빅데이터 마이닝에서 활용되는 기법 중

신경망은 수많은 예측과 클러스터링의 결과를 활용 하는데 많이 사용되는 기법이며, 신경망 모델이 어떻게 처리되는지에 대한 연구보다 결과에 대한 연구가 중요할 경우 자주 활용되는 머신러닝 기법이다. 신경망 알고리즘 중 인공신경망(Artificial Neural Network)은 생물의 신경을 모방한 통계학적 표현 기법이며 머신러닝의 지도학습 알고리즘이다. 사람의 뇌

※ Corresponding Author : Chun Bo Sim, Address: (540-742) Maegok-Dong, Suncheon-si Jeollanam-do, Korea, TEL : +82-61-750-3834, FAX : +82-61-750-3830, E-mail : cbsim@sunchon.ac.kr

Receipt date : May 21, 2018, Revision date : Jun. 5, 2018  
Approval date : Jun. 11, 2018

<sup>†</sup> Dept. of Multimedia Eng., Sunchon National University (E-mail : iam1710@hanmail.net)

<sup>\*\*</sup> Dept. of Computer Eng., Sunchon National University (E-mail : seaghost@sunchon.ac.kr)

<sup>\*\*\*</sup> Major of Game Animation Engineering, Dong-Eui University (E-mail : kimchee@deu.ac.kr)

<sup>\*\*\*\*</sup> Dept. of Library & Information Science, Jeon-ju University (E-mail : gsyoun@jj.ac.kr)

<sup>\*\*\*\*\*</sup> Dept. of Multimedia Eng., Sunchon National University

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. 2017R1D1A3B03035379). And this paper has been written with the support of Jeollanam-do(2017 R&D supporting program operated by Jeonnam Technopark).

는 뉴런(Neuron)이라는 구조로 구성되어 있고 학습을 통해 인지나 패턴 인식 등 다양한 기능들을 처리한다. 생물학적으로 신경망은 신경 세포간의 결합으로 구성되어 있다. 세포들 사이에서는 생물학적인 뉴런(Neuron)정보의 전달이 이루어진다. 신경세포는 정보의 입력을 담당하는 수상돌기, 정보를 합하여 연산을 수행하는 세포체, 정보를 전달 및 출력을 담당하는 축색의 3가지 요소로 구성되어 있다. 정보를 합하여 연산을 수행하는 세포체 내부는 입력의 합이 일정 부분에 도달하면 활동전위를 발생시키는 특징을 가지고 있다. 신경세포들은 시냅스를 통해 연결되고 신경세포들의 연결 세기에 따라 정보 전달 방법과 규모가 달라진다. 이렇게 연결 세기에 따라 정보 전달의 방식과 형태가 달라지며, 데이터를 예측하고 평가하기 위하여 활용되는 신경망 연구가 꾸준히 진행되고 있는 실정이다. 신경망을 활용한 예측 모델 및 클러스터 모델 적용에 관한 기존 연구 방향은 대체적으로 인공신경망을 적용하였다[1-4].

인공신경망은 적합한 예측 모델을 찾기 위해서 모델의 학습 횟수를 증가하는 문제점이 있다. 그리고 데이터의 국소적 및 전역적 최소값 문제로 데이터 모델 예측 정확률이 낮아지는 문제점이 있다. 이를 보완하기 위해 2000년대 중반부터 연구된 딥러닝의 기술은 인공신경망의 문제점을 해결하였지만 기존의 최소값에 대한 문제점은 지속되었다[5-9]. 특히 클러스터링 분야에서 역전파 알고리즘은 과적합의 문제점을 보완하였지만, 전체적인 다량의 학습 데이터 확보가 선결조건으로 부각되었다. 또한 전체 데이터에 대해 일괄적인 학습과 평가가 이루어져야 하는 문제점도 존재하고 있다. 효율적인 빅데이터 분석을 위해서 입력 데이터의 정확한 결과 분석과 새롭게 입력되는 데이터에 대한 예측이 가능해야 한다[4, 11].

본 논문에서는 클러스터 데이터를 활용하여 단계별 데이터 학습 반응이 아닌 전체적인 형태의 데이터 학습 방법 반응이 가능하도록 신경망을 적용하고 클러스터별 학습을 통해 신경망 모델 선정과 변수의 분석을 진행한다. 이를 위하여 K-means 클러스터링의 초기화 접근 방법을 통해 클러스터링 결과의 반응 변수 분석과 새로운 데이터가 입력될 경우 클러스터링의 예측을 및 예측 적중률을 평가하는 모델을 구축하고자 CALR(Clustering  $k$ -LSTM RNN) 알고리즘을 제안한다.

## 2. 관련 연구

J. W. Lee[5]의 연구에서는 일반 텍스트 데이터들을 분류할 수 있는 문서 분류 기법을 제안하였다. 인공신경망에 입력값을 위해 문서들의 의미론적인 추론 모듈을 적용하여 입력값을 추출하는 방법을 제안하였다. 이를 위해 의미론적 추론 모듈은 온톨로지를 활용하였으며, 동시에 인공신경망의 문제점으로 지적된 수치적인 해석이 아닌 의미론적인 해석을 결합한 기법을 제안하였다. 입력값에 대한 학습 모듈은 역전파 알고리즘을 통해 적용하였으며, 델타 룰에 따라 학습을 진행하였다. 그러나 해당 연구는 인공신경망의 문제점으로 지적된 국소적 최소값과 전역적 최소값의 문제점을 해결하지 못해 학습 반복 횟수가 지속적으로 증가하는 문제점을 포함하고 있다. S. M. Jung[6]의 연구에서는 얼굴 특징 벡터와 음성 특징 벡터를 추출하여 사용자 인증이 가능한 시스템을 신경망 알고리즘 기법으로 제안하였다. 해당 연구에서는 이미지와 음성 특징 벡터 영역을 검출한 후 정규화 과정을 거친다. 그런 다음 유사도를 측정하여 인공신경망에 유사도 측정값을 학습시키는 방안을 제시하였다. 이미지의 경우, 아다부스팅 알고리즘과 PCA, K-means 알고리즘을 활용한 유사도 결과값과 본인 이미지에 대한 학습 데이터 정합율이 높은 결과를 도출하였다. 그러나 음성의 경우 왜곡된 현상 또는 길이에 따른 프레임의 문제로 인하여 정합율이 낮게 측정되었다. 그러나 음성 인식은 분류에 주된 목적을 가지고 있으며, 학습 데이터의 크기와 규모보다는 학습 데이터의 특정 시점을 기준으로 이전과 이후의 학습 데이터 연계를 활용한다면 음성의 인식률을 높일 수 있을 것으로 예상된다. H. W. Jeong[7]의 연구에서는 인공신경망의 문제점을 보완한 딥러닝의 DBN 알고리즘을 활용하여 기업부도예측 모형을 제안하였다. 기존 기업부도예측 모형은 SVM을 적용하여 예측 모형에 대한 정확률을 더 높이는 결과를 나타냈다. 이는 딥러닝 알고리즘의 하나인 DBN을 통해 데이터 분류의 민감도에 대한 범위를 높이고 다양한 데이터 분석할 수 있는 장점을 가진다. 그러나 높은 민감도는 동전의 양면성과 같이 국소적 최소값과 전역적 최소값에 대한 오류율을 높이는 문제점을 내재하고 있다. S. H. Ryu[8]에서는 심층신경망을 기반으로 전력 수요 예측 모델에 대한 연구를 진행하였다. 대용량의 실제 수용가 전력 사용량 정보를 통

해 단기 수요 예측 수행과 결과를 인공지능망과 심층 신경망으로 분석하였다. 해당 연구에서는 대용량의 실제 수요가 데이터에 심층신경망 기법을 처음으로 적용하였으며, 정확한 분석을 위하여 실제 수요와 예측 수요 사이의 오차는 평균 절대 백분비 오차(Mean Absolute Percentage Error, 이하 MAPE)와 평균 제곱근 오차(Root Means Square Error, 이하 RMSE)를 적용하여 예측에 대한 결과를 평가하였다. Giles et al.[9]의 연구에서는 순차적 데이터의 예측을 위해 순환신경망에 대한 연구를 제안하였다. 순환신경망이 클러스터링에 대한 문제에서 좋은 연구 성과 및 결과를 도출하는 것에 착안하였다. 클러스터링 모델 기반 기법인 자기조직지도(Self Organization Map, 이하 SOM)에 활용하여 클러스터링을 실행한 후, 이를 반응 변수로 사용하였다. 순차적 데이터 예측을 목표로 하여 하나의 실수값을 예측값으로 갖는 회귀 분석이 아닌 클러스터링을 위해 순환신경망 연구를 제시했다.

### 3. 제안하는 클러스터별 검증을 위한 RNN 알고리즘 설계

#### 3.1 제안하는 데이터 분류 평가 예측 모델

본 연구에서 제안하는 방법은 K-means 클러스터링을 활용하여 전체 데이터를 분류하고 실제 데이터 집합의 예측 및 검증을 목적으로 RNN 알고리즘을 활용한다. 예측 모델의 신경망은 클러스터링으로 분류된 클러스터별로 신경망을 구축하여 반응 변수 예측과 모델 검증을 진행한다. 그리고 전체적인 학습 반영과 오류 수정이 다양한 층에 전달될 수 있도록 딥러닝 기법 중 하나인 RNN 알고리즘을 활용하여 데이터 훈련 및 평가 모델을 설계한다. Fig. 1은 본 논문에서 제안하는 시스템의 전체 구성도이다[10].

#### 3.2 클러스터별 검증을 위한 RNN 훈련 모델 설계

기존에 연구[9]된 RNN 예측 모델 훈련 구조는 순차적 데이터를 클러스터링할 경우, 훈련의 손실값(이하 Loss Value)을 계산하고 이를 최소화한 후 가중치 값(이하 Weight Value)의 최적 모델을 선택하는 훈련을 연구하였다. 그러나 한 번의 훈련을 위하여 매번 반복마다 입력값의 데이터 크기와 출력값의 데이터의 크기가 정해져 있다는 문제점이 있다. 본

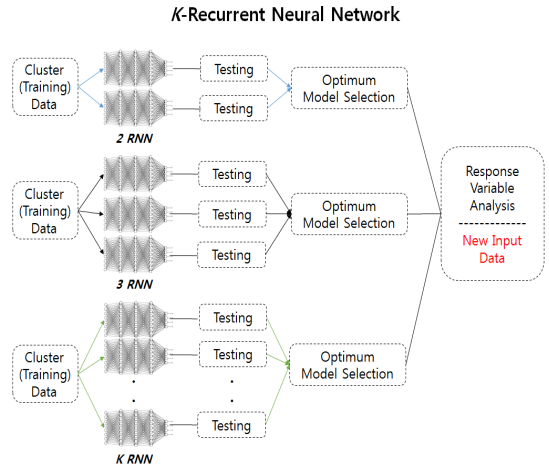


Fig. 1. Overall Structure of proposed K-Recurrent Neural Network.

논문에서는 CkLR의 알고리즘을 제안하면서 2가지의 조건을 기반으로 훈련 구조를 설계한다. 첫 번째로는 훈련 모델에 활용하고자 하는 클러스터별 데이터가 비순차적 데이터로 주어질 때, 훈련 모델의 이전에 존재했던 값들이 예측 모델에 필요하다고 가정하며, 두 번째로는 CkLR 알고리즘이 진행될 경우, 중간에 상태 노드에 대한 초기화는 진행되지 않아야 한다는 것이다. 초기화가 진행되지 않는 훈련 구조에 반영할 경우 현재의 훈련 상태에서 지난 과거의 모든 훈련 정보를 새로운 노드가 모두 알 수 있으며, 다음 노드의 예측 시 반영할 수 있다는 장점을 가지고 있다. Fig. 2는 본 논문에서 제안하는 CkLR 훈련 모델의 일부 구조이다.

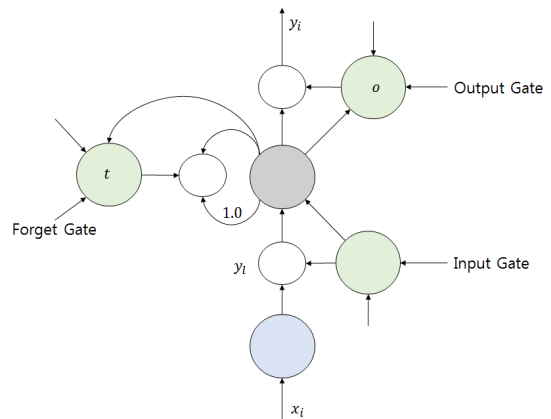


Fig. 2. Structure of CkLR.

제안하는 CkLR의 훈련 모델 알고리즘은 아래 4가지의 필수조건을 포함한다. 첫 번째로는 벡터간의 정보 전달을 위하여 요소별 곱(Element wise product)  $\odot$ ,  $i$ 번째 항(이하 Term)의 입력은  $x_i$ , 학습을 통해 결과값이 결정되는 가중치 행렬은  $W$ 이며, 바이어스 항(Bias Term)은  $b$ 이며, 시그모이드(Sigmoid)함수는  $\sigma$ 로 구분한다. 두 번째로는 Term의 입력 정보에 대하여 사용 여부를 결정하는 Input-Gate는 식 (1)과 같이  $t$ 이며, Term에서 Unit의 상태에 대하여 출력 여부를 결정하는 Output-Gate는 식 (2)와 같이  $o$ 로 구분한다. 세 번째로는 LSTM-Cell로 직접적으로 연관된 Unit의 현재 Term 상태는 식 (3)과 같이  $c$ 이며, Forget-Gate로 LSTM-Cell이 Unit의 이전 Term의 상태를 기억하여 현재의 Sequence에 사용 여부를 결정하는 것은 식 (4)와 같이  $f$ 로 구분한다. 마지막으로 은닉층(Hidden Layer)의 출력은  $h$ 이며, 활성화 함수(e.g. Sigmoid, tanh)은  $act$ 이며, 출력층(Output Layer)의 출력은 식 (5)와 같이  $y$ 로 구분한다.

$$t_i = \sigma(W_{tx}x_i + W_{th}h_{i-1} + W_{tc}c_{i-1} + b_t) \quad (1)$$

$$o_i = \sigma(W_{ox}x_i + W_{oh}h_{i-1} + W_{oc}c_{i-1} + b_o) \quad (2)$$

$$c_i = f_i \odot c_{i-1} + t_i \odot \tanh(W_{ix}x_i + W_{ih}h_{i-1} + W_{ic}c_{i-1} + b_i) \quad (3)$$

$$f_i = \sigma(W_{fx}x_i + W_{fh}h_{i-1} + W_{fc}c_{i-1} + b_f) \quad (4)$$

$$y_i = g(W_{yh}h_i + b_y) \quad (5)$$

### 3.3 클러스터별 검증을 위한 RNN 평가 모델 설계

클러스터링을 통해 출력되는 클러스터링의 결과값이 비순차적인 구조를 가지고 있기 때문에 입력 데이터의 길이가 특정값으로 종속되지 않는다. 따라서 Loss Value를 어느 단계에서 얻을 것인지를 설계해야 한다. 본 논문에서는 알고리즘을 진행하는데 있어 최적화를 진행하는 입력과 출력의 형태를 하나의 단계로 구분하고 입력 데이터에 값을 할당하는 Single Step 모형으로 연구를 진행한다. 훈련 모델은 클러스터에 포함된 데이터를 활용하고 결과를 반영하여 예측을 측정하는 것으로 정의한다. 매번 반복하는 시점에서 데이터의 길이만큼 반복 횟수를 지정하고 Loss Value를 측정한다. 그리고 그 값을 최소화하는 것으로 훈련 모델의 Weight Value를 최적화한다. Fig. 3은 Loss Value를 적용하기 위한 Single Step CkLR의 훈련 모델을 도식화한 것이다.

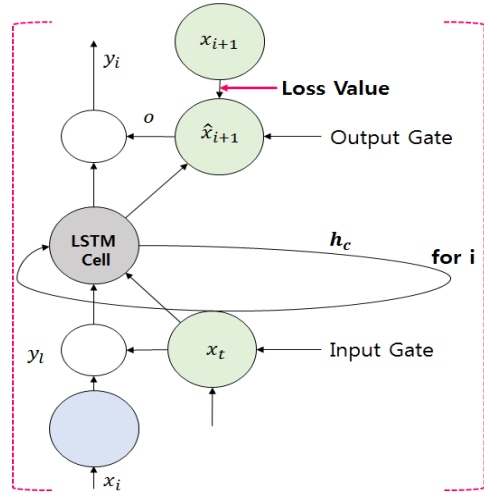


Fig. 3. Learning Method of Single Step CkLR.

Layer 구조는 신경망 알고리즘을 적용하기 위한 Layer 수를 어떻게 정할 것인지를 결정하는 것이며, 훈련 모델의 구조 설계 시 LSTM Cell의 Layer 수를 고려해야 한다. 본 연구에서 제안한 알고리즘은 클러스터링의 결과에 따라 클러스터의 수 만큼 신경망 구조를 적용하기로 정의하지만, 입력 데이터와 클러스터의 수가 유동적인 관계로 본 논문에서는 훈련 모델의 Layer를 Single Layer로 한정하여 연구를 진행한다.

Fig. 4는 제안한 CkLR의 전체 알고리즘이다. 통합된 예측 모델 신경망은 전체 데이터 집합을 그대로 사용하여 반응 변수를 예측하지만 본 논문에서 제안하는 CkLR은 입력되는 전체 데이터를 K-means 클러스터링을 진행하고 각 클러스터에 배정된 LSTM-RNN이 해당 클러스터에 배정된 데이터를 예측하게 된다. 알고리즘이 정립된 이후에 새로운 입력 데이터가 존재하면 클러스터 배정 여부를 진행한 후 해당 클러스터에 배정된 CkLR을 통해 데이터의 반응 변수를 예측률을 평가하게 된다.

## 4. 제안하는 알고리즘의 성능평가

### 4.1 성능평가 데이터 및 범위

성능평가를 위해 적용되는 데이터 집합은 Image Segmentation[11]의 데이터를 활용하였으며 데이터 집합의 데이터 개수는 2310개의 데이터로 Table 1과 같이 구성되어 있다. 학습 데이터, 테스트 데이터, 평

- ① Check Cluster Number  $q$  identified with the entered Data Set  $X$  and K-means Clustering.
  - ①-① Separate it into Training Data  $X_{train}$  and Test Data  $X_{test}$ .
  - ①-② Redefine the defined data from the time of the previous stage.
- ② Do learning for the weight values of LSTM.
- ③ Change the state node values of LSTM by using Training Data  $X_{train}$  again and, at the same time, update  $h_e$ , the information of the Previous State Node  $x_{t-1}$ .
- ④ Do learning with  $x \leftarrow X_{test}[i]$ , the first input value to test the first test data.
- ⑤ Obtain the output of prediction and response variables for data by repeating learning to the length of the entered clusters.

Fig. 4. Proposed C/RLR Algorithm.

Table 1. Data Classification of Image Segmentation Data Set

Class	Learning Data	Test Data	Evaluation Data	Total
Grass	198	99	33	330
Sky	198	99	33	330
Cement	198	99	33	330
Foliage	198	99	33	330
Window	198	99	33	330
Brickface	198	99	33	330
Path	198	99	33	330
Total	1,188	594	198	1,980

가 데이터의 비율은 6:3:1로 구분하여 성능평가를 진행하였다.

클러스터링된 반응 변수를 분석하고 새로운 변수를 예측하는 방법으로 제안하는 C/RLR 알고리즘은 2가지 척도를 기준으로 알고리즘에 대한 성능을 확인한다[12]. 첫째, 데이터를 클러스터별로 분류함으로써 클러스터링된 반응변수 간의 규칙이나 패턴 분석이 용이하도록 클러스터링  $k$ -LSTM 상태 노드에 따른 훈련 데이터 및 테스트 데이터 적합성에 대한 평가이며 둘째, 클러스터링  $k$ -LSTM 상태 노드에 따른 클러스터별 Loss Function을 측정한다.

#### 4.2 클러스터별 데이터 적합성 평가

Image Segmentation의 19차원인 다차원 데이터 집합을 본 논문에서 제안하는 PKM[13]에 적용하여 차원 축소 및 클러스터링을 진행한 결과를 바탕으로 클러스터 수를 측정하였고, 주성분 분석을 통한 차원 축소의 범위(전체 데이터의 80%이상을 설명할 수 있는 범위) 결과는 6차원으로 출력되었다. 이러한 결과

는 주성분 1부터 주성분 6까지 클러스터 K값의 관측 범위로 선택된다. 관측범 위내의 모든 성분에 대하여 비 유사도 값은 주성분 6이 가장 높게 측정되었으며, 6차원일 경우 최적의 클러스터 수로 결정한다. 클러스터별 신경망의 정확성을 측정하기 위하여 학습 데이터 및 테스트 데이터에 대한 적합성을 측정한다. 제안하는 클러스터별 신경망에 대한 데이터 적합성을 측정하는 방법은 단일 신경망과 클러스터별 신경망의 데이터 적합성을 각각 측정하여 비교 평가하는 방법으로 진행한다. 실험에서 신경망 은닉층은 3개 층으로 고정하고, 학습율은 0.0007, 모멘텀은 0.0002, 학습 데이터의 학습 횟수는 20,000번을 반복했다.

Table 2는 19차원의 다차원 데이터를 K-means 알고리즘을 적용하여 최적 클러스터로 구분하였고 클러스터별 신경망 테스트 데이터 적중률을 측정한 결과이다. 각 Class별 테스트 데이터 적중률이며, Cluster 1은 Path Class(96.64%)와 Window Class(93.84%)가 높은 데이터 적중률을 보였다. Cluster 2는 Foliage Class(96.98%)의 적중률이 가장 높았으며, Cluster 3은 Brickface Class(93.15%)와 Cement Class(95.32%), Cluster 5는 Grass Class(94.56%), Cluster 6은 Sky Class(93.33%)가 높은 데이터 적중률을 보였다. Class별 테스트 데이터 적중률은 전체 테스트 데이터 적중률의 평균인 92% 넘어서는 적중률로 제안하는 Class별 학습 신경망의 특화된 구조를 가진 장점과 부합한다.

#### 4.3 클러스터별 Loss Function 성능평가

C/RLR에 따른 데이터 분석과 예측 결과에 대한 정확성을 확인하기 위해 클러스터별 개체를 기반으로 반응 변수 예측 결과를 비교하는 클러스터별 손실함

Table 2. Test Data Hit Rate by Cluster(k=6)

Item	Brickface	Sky	Foliage	Cement	Window	Path	Grass
Cluster 1	92.46	90.04	91.64	93.46	93.84	96.64	91.18
Cluster 2	89.46	90.18	96.98	92.44	91.69	93.47	92.46
Cluster 3	93.15	88.64	91.64	95.32	92.14	91.23	92.77
Cluster 4	91.64	91.11	90.48	90.22	91.88	90.86	92.43
Cluster 5	93.11	90.61	92.47	93.11	93.01	92.73	94.56
Cluster 6	90.45	93.33	88.91	90.09	90.46	90.66	91.45

수(Loss Function) 측정이 필요하다. 손실함수는 미리 정의된 분석 결과를 기준으로 CMLR을 통한 데이터 분석 및 예측 결과의 차이를 측정하는 함수이다. 성능 실험에서는 클러스터별 훈련 데이터 및 테스트 데이터를 기반으로 RMSE를 측정한다. RMSE는 훈련 데이터에 적용할 경우 적합도를 의미하며 테스트 데이터에 적용할 경우 예측도를 의미한다. 결과 수치가 낮을수록 적합도와 예측도의 성능이 우수하다. RMSE 측정식은 식 (6)과 같다.

$$RMSE = \sqrt{\frac{\sum(x_i - x)^2}{N}} \quad (6)$$

$x_i$  : Actual Value

$x$  : Prediction Value

본 논문에서 제안한 Clustering  $k$ - LSTM RNN 형태는 단일 스텝과 단일층 모형으로 알고리즘을 제안하였다. 단일 스텝은 1개 층의 순환신경망으로 훈련시켰으며, LSTM 상태 노드 수는 1, 5, 10으로 5단계 차이를 적용했으며, 학습률은 0.0007을 적용하여 성능평가를 진행한다. 학습률을 변경하는 목적은 신경망의 가중치 값을 학습할 경우 가중치의 부호가 급하게 변하는 것을 막거나 변화하는 범위를 줄이기 위해서 학습률을 조절하는 것이다. Table 3은 Brick-

face Class의 데이터를 활용하여 분류된 각 클러스터별로 LSTM 상태 노드수를 적용하고 테스트 데이터에 대한 RMSE를 측정한 결과이다. 3번째 클러스터에서 LSTM 상태 노드 수가 10일 때 테스트 데이터의 예측도가 가장 우수하다. 평균적으로 학습률이 0.007일 경우 Cluster 3의 예측도가 가장 좋은 결과를 나타낸다.

## 5. 결 론

기존 연구에서 진행된 신경망 역전파 알고리즘은 과적합의 문제점을 보완하였지만, 전체적인 다량의 학습 데이터 확보가 선결조건으로 부각되었다. 또한 전체 데이터에 대해 일괄적인 학습과 평가가 이루어져야 하는 문제점도 존재하고 있다. 이에 본 논문에서는 클러스터 데이터를 활용하여 단계별 데이터 학습 반영이 아닌 전체적인 형태의 데이터 학습 방법 반영이 가능하도록 신경망을 적용하고 클러스터별 학습을 통해 반응 변수 분석 및 검증을 위해 CMLR 알고리즘을 제안했다. CMLR 알고리즘은 학습 데이터의 학습 과적합 문제를 해결하기 위하여 주성분 분석 및 예상되는 이상점 활용을 통해 클러스터링된 클러스터별로 신경망을 적용하여 최적의 예측 및 분

Table 3. Cluster RMSE by LSTM State Node (Learning Ratio=0.0007)

Cluster	Learning Ratio	LSTM State Node		
		1	5	10
1	0.0007	0.8846	1.2455	1.2452
2	0.0007	0.9875	0.9655	1.1253
3	0.0007	1.2481	0.6452	0.0544
4	0.0007	0.8974	0.2345	0.4322
5	0.0007	1.1024	0.9855	0.6488
6	0.0007	0.9878	1.1245	0.5444
Average		1.0180	0.8668	0.6751

석 모델을 선택할 수 있는 알고리즘이다.

단 한번의 훈련을 위하여 매번 반복마다 입력값의 데이터 크기와 출력값의 데이터의 크기가 정해져야 하는 문제점을 보완하기 위하여 LSTM 순환신경망 기반의 CCLR 알고리즘을 제안했다. 제안된 신경망 예측 모델 훈련 구조는 순차적 데이터를 클러스터링 할 경우, 훈련의 손실값을 계산하고 이를 최소화한 후 가중치가 반영된 최적의 모델을 선택하는 훈련 방식이다. 성능 실험 결과 기존 하나의 전체 신경망 구조보다는 개별 신경망을 통한 예측율과 적중률이 더 우수하다는 것을 확인하였다.

## REFERENCE

- [1] D.H. Shin, K.H. Choi, and C.B. Kim, "Deep Learning Model for Prediction Rate Improvement of Stock Price Using RNN and LSTM," *Journal of Korea Information Technology Association*, Vol. 15, No. 10, pp. 9-16, 2017.
- [2] D.H. Shim, "Inductive Learning Using Theory-Refinement Knowledge-Based Artificial Neural Network," *Journal of Korea Multimedia Society*, Vol. 4, No. 3, pp. 280-285, 2001.
- [3] S.B. Park and D.S. Yoo, "Multimedia Database Management Issues," *Journal of Korea Multimedia Society*, Vol. 21, No. 3, pp. 391-399, 2018.
- [4] S.H. Jung, J.C. Kim, and C.B. Sim, "A Novel Data Prediction Model Using Data Weights and Neural Network Based on R for Meaning Analysis between Data," *Journal of Korea Multimedia Society*, Vol. 18, No. 4, pp. 524-532, 2015.
- [5] J.W. Lee, *An Automated Text Classification Method Using Multi Feature Extraction Module Based on Neural Network*, Master's Thesis of Hanbat National University, 2016.
- [6] S.M. Jung, *Study on the Integrated System of Face Recognition and Speaker Verification Using Neural Network*, Master's Thesis of Hongik University, 2015.
- [7] H.W. Jung, *Bankruptcy Prediction Based on Deep Learning Algorithm*, Master's Thesis of Hanyang University, 2016.
- [8] S.H. Yoo, *DNN Based Customer Electric Load Forecasting*, Master's Thesis of Sogang University, 2016.
- [9] C. Lee Giles, Steve Lawrence, and A.C. Tsoi, "Noisy Time Series Prediction Using Recurrent Neural Networks and Grammatical Inference," *Journal of Machine Learning*, Vol. 44, No. 1, pp. 161-183, 2001.
- [10] S.H. Jung, K.J. Kim, J.C. Kim, C.Y. Kim, and C.B. Sim, "A Study on Recurrent Neural Network for Prediction and Accuracy Improvement of Data Classification Evaluation," *Proceeding of the Fall Conference of the Korea Multimedia Society*, pp. 49-51, 2017.
- [11] Machine Learning Repository(2018). <http://archive.ics.uci.edu/ml> (accessed Feb., 15, 2018).
- [12] S.H. Jung, *A Novel on Hybrid Machine Learning Method Based on Big Data Mining*, Doctor's Thesis of Suncheon National University, 2017.
- [13] S.H. Jung, K.J. Kim, E.C. Lim, and C.B. Sim, "A Novel on Automatic K Value for Efficiency Improvement of K-means Clustering," *Proceeding of International Conference on International Conference on Future Information Technology, International Conference on Multimedia and Ubiquitous Engineering*, Vol. 448, pp. 181-186, 2017.



정 세 훈

2010년 2월 순천대학교 멀티미디어공학과 공학사  
2012년 2월 순천대학교 멀티미디어공학과 공학석사  
2017년 2월 순천대학교 멀티미디어공학과 공학박사

2015년 1월 ~ 현재 광양만권 SW융합 연구소 팀장  
관심분야: 빅데이터 처리 및 확률 분석, 데이터마이닝, 감성분석, 객체지향 모델링



유 강 수

1991년 2월 전북대학교 컴퓨터공학과 공학사  
1994년 2월 전북대학교 대학원 컴퓨터공학과 공학석사  
2005년 2월 전북대학교 대학원 영상공학과 공학박사

2006년 9월 ~ 현재 전주대학교 교수  
관심분야: 영상처리, 컴퓨터비전, 멀티미디어시스템



김 종 찬

2000년 2월 순천대학교 컴퓨터공학과 공학사  
2002년 2월 순천대학교 컴퓨터공학과 이학석사  
2007년 8월 순천대학교 컴퓨터공학과 이학박사

2012년 9월 ~ 2013년 8월 서울대학교 자동화시스템연구소 선임연구원  
관심분야: 영상처리, 컴퓨터 그래픽스, 디지털클로딩, 증강현실, 빅데이터, 얼굴인식, 게임, HCI



심 춘 보

1996년 2월 전북대학교 컴퓨터공학과 공학사  
1998년 2월 전북대학교 컴퓨터공학과 공학석사  
2003년 2월 전북대학교 컴퓨터공학과 공학박사

2005년 3월 ~ 현재 순천대학교 정보통신·멀티미디어공학부 교수  
관심분야: 빅 데이터 시스템, 머신러닝, IoT/IoE 플랫폼, 멀티미디어 시스템



김 치 용

1991년 2월 인제대학교 물리학과 이학사  
2000년 2월 인제대학교 전산물리학과 이학박사  
2000년 ~ 2006년 부산정보대학 정보통신계열 및 동서대학교 디지털 디자인학부 조교수

2007년 영국 옥스퍼드대학교 Harris Manchester College, Visiting Fellow  
2012년 서울대학교 자동화 시스템 공동연구소 객원교수  
2006년 3월 ~ 현재 동의대학교 디지털 콘텐츠 게임 애니메이션공학부 교수  
관심분야: 3D Animation, Multimedia Design, Chaos & Fractal Design, VR Contents Design, 게임캐릭터디자인, 가상 피팅시스템, Computational Simulation