https://doi.org/10.5392/JKCA.2018.18.11.416

# LSTM 네트워크를 활용한 농산물 가격 예측 모델

## A Prediction Model for Agricultural Products Price with LSTM Network

신성호*, 이미경*, 송사광**
한국과학기술정보연구원 연구데이터플랫폼센터*,
한국과학기술정보연구원 연구데이터플랫폼센터/과학기술연합대학원대학교 빅데이터과학과**

Sungho Shin(maximus74@kisti.re.kr)*, Mikyoung Lee(jerryis@kisti.re.kr)*,
Sa-kwang Song(esmallj@kisti.re.kr)**

**요약**

태풍, 홍수는 우리에게 빈번하게 닥치는 자연 재해이며, 이와 같은 자연 재해로부터 오는 피해는 사전에 예측되어 대응책이 마련될 필요가 있다. 자연 재해로부터 야기되는 피해에는 건물의 붕괴, 인명 피해, 논/밭의 유실 등 주로 직접적인 피해가 많지만, 소비자 물가 상승과 같은 간접적인 영향에도 관심을 가져야 한다. 태풍, 홍수의 피해로부터 영향을 받는 대표적인 소비재 상품은 농산물이다. 갑작스럽고 강력한 태풍은 많은 비를 동반하면서 농작물에 피해를 주고, 농산물의 가격을 상승시킨다. 본 연구에서는 딥러닝 알고리즘을 활용하여 태풍과 같은 자연 재해가 농산물 가격에 미치는 영향을 예측한다. 우리는 데이터 확보가 가능한 쌀, 양파, 대파, 애호박, 시금치 등을 가격 예측 대상으로 했고, 농산물 가격에 영향을 미치는 변수 데이터들로 학습 모델을 만들고, 그 학습 모델이 농산물 가격을 예측하는 연구를 진행하였다. 연구 결과, 모델에 의한 예측 가격과 실제 가격의 차이를 나타내는 RMSE가 0.069 수준이며, 농산물 가격을 비교적 잘 설명하는 것으로 해석된다. 정확한 농산물 가격 예측은 정부의 농산물 공급 규모 조절 등 자연 재해 대응을 위한 정부의 노력에 활용될 수 있을 것이다.

■ 중심어 : | 자연재해 | 농산물 | 가격 | 예측 | 딥러닝 | LSTM | RMSE | 날씨 |

**Abstract**

Typhoons and floods are natural disasters that occur frequently, and the damage resulting from these disasters must be in advance predicted to establish appropriate responses. Direct damages such as building collapse, human casualties, and loss of farms and fields have more attention from people than indirect damages such as increase of consumer prices. But indirect damages also need to be considered for living. The agricultural products are typical consumer items affected by typhoons and floods. Sudden, powerful typhoons are mostly accompanied by heavy rains and damage agricultural products; this increases the retail price of such products. This study analyzes the influence of natural disasters on the price of agricultural products by using a deep learning algorithm. We decided rice, onion, green onion, spinach, and zucchini as target agricultural products, and used data on variables that influence the price of agricultural products to create a model that predicts the price of agricultural products. The result shows that the model's accuracy was about 0.069 measured by RMSE, which means that it could explain the changes in agricultural product prices. The accurate prediction on the price of agricultural products can be utilized by the government to respond natural disasters by controling amount of supplying agricultural products.

■ keyword : | Natural Disaster | Agricultural Product | Price | Prediction | Deep Learning | LSTM | RMSE | Weather |

* 본 연구는 한국과학기술정보연구원의 연구비를 지원받아 수행되었습니다.
접수일자 : 2018년 10월 24일
심사완료일 : 2018년 11월 09일
수정일자 : 2018년 11월 09일
교신저자 : 신성호, e-mail : maximus74@kisti.re.kr

## I. 서 론

자연 재해, 질병 확산 등과 같이 인간의 안전을 위협하는 사회 현안 문제들에 대한 연구는 지속적인 주목을 받으며 꾸준히 진행되고 있다. 이런 이슈들은 발생 확률이 높지 않지만, 한 번의 발생으로 많은 생명과 재산 피해를 유발시킨다. 때문에 기상은 우리 생활에 밀접한 영향을 주는 요소 중 하나이다. 기상에 따라 여행 지역이 결정되기도 하고, 심지어 우리 기분까지 좌우되기도 한다. 농사를 짓는 농부들에게 기상정보는 반드시 필요한 정보이다[1]. 기업 활동도 기상에 영향을 많이 받는다. 많은 크고 작은 기업들도 기업 경영에 기상 정보를 활용한다.

작은 날씨 변화와는 다르게 갑작스럽고 극심한 날씨 변화는 자연 재해의 형태로 우리에게 큰 피해를 주기도 한다. 태풍 또는 홍수와 같은 자연 재해로부터 야기되는 직접적인 피해로는 건물의 붕괴, 인명 피해, 논/밭의 유실 등 여러 가지가 있다. 소비자 물가의 변화는 직접적이지는 않지만, 우리의 삶에 적지 않은 영향을 미친다. 자연 재해로 인해 농산물이나 과일 가격이 많이 올랐다는 불만의 말들을 자주 들을 수 있다. 특히, 우리나라는 7월에서 10월까지 태풍 피해가 자주 발생하는데, 이 기간 중에 추석이라는 명절이 있다. 추석이라는 명절이 다가오면 대부분의 가정에서 명절 음식에 들어가는 재료들을 구매하기 때문에 소비자 물가가 많이 올라간다. 이 기간 태풍까지 오게 되면 소비자 물가는 상상할 수 없을 정도로 뛰어 오를 수 있다.

이러한 자연 재해의 영향을 예상해서 정부에서는 물가 조절을 위해 농산물이나 과일과 같은 원재료들을 시중에 더 공급하는 것으로 물가를 안정시키기 위해 노력한다. 하지만, 구체적으로 어느 정도 오를지, 또는 오를지 안 오를지에 대한 정확한 분석 없이 과거의 경험에 기반해서 의사결정이 이루어지고 있다. 정확한 농산물 가격 예측 모델의 확보는 자연 재해 시 농산물 가격의 예측, 정부 차원의 농산물 공급 규모, 수입량 등 자연 재해 대비를 위한 정부의 노력에 유용하게 활용될 수 있을 것이다.

본 연구에서는 인공신경망의 진보된 모습이자 현재 가장 우수한 성능을 보이는 예측 기술 중 하나인 딥러닝 기술을 이용해서, 자연재해 발생 시 농산물의 가격이 어떻게 변화하는지 예측하는 모델을 만들고, 그 모델의 정확도를 실험을 통해 제시한다. 우리는 데이터 확보가 가능한 쌀, 양파, 대파, 애호박, 시금치 등을 예측 대상으로 했고 관련 데이터를 수집하여 딥러닝 알고리즘을 적용하였다.

연구를 위해, 3장에서는 예측 모델에 활용된 변수 및 데이터에 대한 설명과 예측 모델의 생성을 위해 사용된 딥러닝 알고리즘에 대해서 소개한다. 데이터 수집 및 전처리 등 딥러닝을 활용하기 위한 데이터 구축에 대한 내용을 4.1장에서 기술하고, 학습 모델을 구축하기 위해 사용한 알고리즘, 하이퍼 파라미터 세팅, 학습 환경은 4.2장에서 기술한다. 5장에서는 2000년부터 17년치 데이터를 활용한 학습 및 테스트 결과에 대해서 기술한다.

## II. 관련 연구

본 연구는 쌀, 시금치, 애호박, 대파, 양파 등 5가지 농산물을 대상으로 주요 대도시에서 거래되는 소매 가격을 예측하고, 태풍과 같은 자연재해 발생 시 농산물 가격 변화를 예측하는 것이다. 이전에 이루어졌던 농산물들에 대한 시계열 분석은 전통적인 방법들이 주로 활용되어져 왔고, 최근의 인공신경망이나 딥러닝을 적용한 연구들은 많이 찾아볼 수 없다. 인공신경망이나 딥러닝 기술이 이미지 인식이나 언어처리에 주로 활용되어 왔기 때문일 것이다. 본 장에서는 먼저 가격 예측에 대한 기존 연구들을 조사 및 분석하였다. 가격 예측 대상은 주로 주식이나 석유 가격 예측에 대한 연구들을 많이 찾아볼 수 있었다. 연구에 적용하는 방법 또는 기술이 딥러닝이기 때문에, 딥러닝에 대한 기본적인 문헌 연구와 특정 알고리즘에 대한 분석도 진행하였다. 마지막으로, 딥러닝을 활용한 농산물 가격 예측 모델의 학습을 위해 어떤 데이터가 필요한지를 도출하기 위해 농산물 가격 예측 연구들에서 활용되었거나 제시된 변수들에 대해서 분석하였다.

### 1. 가격 예측 연구

가격 변화라는 것은 일정한 시간 상에서 독립 변수들의 움직임에 따라 가격이라는 종속 변수가 어떤 변화를 가지느냐는 것이다. 가격 예측에는 전통적으로 시계열 분석 기법들이 많이 사용되고 있으며, 주로 주식 가격 예측과 석유 가격 예측에 대한 연구들이 많이 진행되어 왔다. 농산물 가격 예측은 상대적으로 활발하게 이루어지지는 않았다. 딥러닝을 활용한 농산물 가격 예측 연구들은 더더욱 찾아보기 어렵다. 인공신경망 방법론을 적용한 연구들로 확장하여도 연구 사례들은 많지 않다.

Rather는 주식의 수익률 예측을 위해 Recurrent Neural Network(RNN)을 활용하되 선형 모델(linear model)과 비선형 모델(non-linear model)을 혼합한 하이브리드 모델을 제시하였다[2]. N개의 선형 모델과 M개의 비선형 모델을 각각 생성한 후 각각의 모델에 가중치를 부여하여 하이브리드 모델을 최적화 하였다. Chiroma는 West Texas Intermediate(WTI) 원유 가격 예측 연구를 수행하였고, Genetic Algorithm and Neural Network(GA-NN) 기반으로 하는 접근법을 제안하였다[3]. 실험 결과, 제안된 GA-NN 접근법이 예측 정확도와 계산 효율면에서 기존 알고리즘보다 우수함을 제시했다. 연구들에서 볼 수 있듯이, 최근의 가격 예측 연구들에서는 신경 회로망(Neural Network)을 활용하거나 딥러닝 알고리즘들이 가격 예측에 조금씩 활용되고 있고, 성능 측면에서 시계열 예측 기법과 같은 기존 연구 모델들을 뛰어넘고 있음을 알 수 있다.

### 2. 농산물 가격 예측 연구

Zhang은 Wavelet Neural Network(WNN)을 활용하여 토마토 소매가격을 예측하는 연구를 하였다[4]. 2013년 1월 1일부터 2013년 12월 31일까지 중국의 허베이성의 10개 지역으로부터 토마토 소매 가격 데이터를 수집하였고, 토마토 가격을 예측하는 시계열 모델을 만들어 정확도를 테스트하였다. 결과는 모델 예측 오차율이 0.01 미만이며 예측 값과 실제 값의 상관관계(R2)가 0.908로 토마토의 가격 변동을 비교적 정확하게 예측하였다. 이 연구에서 사용된 변수들은 토마토 가격 외에 기상, 수요/공급량, 토마토 이외의 제품의 시장 가격 변동 등이다.

또 다른 연구는 중국의 농산물(오이, 시금치) 가격 단기 예측 시스템 구축에 대한 것이다[5]. 시스템에서는 여러 종류의 야채의 가격 변동 요인의 유효 정도를 측정하기 위해 다변수 회귀 모델을 구축하였다. 가격 변동 요인으로 유가, 날씨 변화, 계절적 변동 효과가 고려되었다.

임지연은 오이와 호박에 대해서 중기 가격 예측 연구를 수행하였다[6]. 이를 위해 중기선형예측모형을 사용하여 시계열 자료 모형을 설정하였다. 연구를 위해 활용된 변수는 시장 반입 물량, 전국 생산량, 시장 시세, 일일 가격 표준편차 등이다.

남국현은 양파 출하시기의 도매가격을 예측하는 모형을 개발했다[7]. 자기회귀시차 모형을 사용해서 월별 도매가격을 예측하였고, 과거 도매가격, 재배면적, 전년도 도매가격, 농가소득, 농가총수입 등의 데이터를 활용하였다.

배경태는 은닉층(Hidden Layer)이 1개인 기본적인 인공신경망 기법을 이용하여 오이의 가격을 예측하는 연구를 수행하였다[8]. 예측에 활용된 변수는 과거 오이 가격, 30cm정시지중온도, 0.5m정시기온, 정시조도 등 16개 기상관측 데이터이다.

농산물 가격 예측과 관련된 선행 연구를 통해 기상, 수요/공급량, 시장 가격 변동, 유가, 계절적 변동 효과,

**표 1. 농산물 가격 예측 관련 변수**

| 연구 | 농산물 품목 | 변수 |
|---|---|---|
| J. H. Zhang, et al. | 토마토 | 1. 기상<br>2. 수요/공급량<br>3. 시장 가격 변동 |
| C. Wang et al. | 오이, 시금치 | 1. 유가<br>2. 날씨 변화<br>3. 계절적 변동 효과 |
| 임지연 등 | 오이, 애호박 | 1. 시장 반입 물량<br>2. 전국 생산량<br>3. 시장 시세<br>4. 일일 가격 표준편차 |
| 남국현, 최영찬 | 양파 | 1. 과거 도매가격<br>2. 생산량<br>3. 재배면적(농가소득, 농가총수입) |
| 배경태 등 | 오이 | 1. 과거 오이 가격<br>2. 농업기상관측 데이터(30CM정시지중온도, 0.5M정시기온, 정시조도, 10CM정시토양수분 등 16개 관측 데이터) |

전국 생산량, 일일 가격 표준편차, 재배면적, 농가소득, 농가총수입 등의 변수들이 사용된 것을 알 수 있었다. 기상 또는 날씨 정보가 공통적으로 중요한 변수로 사용되었고, 그 외에 가격, 생산량/재배면적, 농가소득 등이 변수로 제시된 것을 볼 수 있다[표 1].

### 3. 딥러닝 연구

가격 예측을 포함하여 향후 일어날 현상에 대한 예측은 회귀분석이나 시계열 분석 등의 방법들을 통해서 수행될 수 있다. 기존에는 회귀분석이나 시계열 분석들이 Matlab과 같은 통계 도구들이나 R 프로그래밍을 통해 수행되어져 왔다. 기존 도구들은 수학 및 통계 공식을 바탕으로, 전체 모집단 중 현상을 설명할 수 있는 샘플 집단의 데이터를 활용하여 예측을 한다. 이에 반해, 딥러닝 기반의 회귀분석이나 시계열 분석은 인간의 인지 및 사고 방식과 유사한 형태로 학습되어진 모델을 통해서 현상을 예측한다. 기존의 방법들은 적당량의 데이터만 있으면 예측이 가능하기 때문에, 분석을 위한 대량의 컴퓨팅 자원이나 인프라를 필요로 하지 않는다. 반면, 딥러닝은 대량의 데이터를 필요로 하며 데이터양에 비례하여 컴퓨팅 자원과 학습을 위한 시간도 필요하다. 이런 단점에도 불구하고 최근 딥러닝이 주목받는 이유는 컴퓨팅 자원 및 인프라의 기술이 어느 정도 발전했고, 학습을 위한 데이터도 폭발적으로 증가하고 있기 때문이다. 예측의 정확도 면에서도 기존의 방법들보다 우수하기 때문에 기계번역, 음성인식, 사물인식 등 다양한 분야에서 활용되고 있다.

딥러닝은 패턴 인식 문제 또는 자질 학습을 위한 많은 수의 뉴런을 갖는 모델을 구성하는 기계 학습 기법이다. 이는 실제 인간의 두뇌가 뉴런 사이의 깊은 연결 구조를 가지고 있다는 점과 유사하다는 시각에서 보다 진보된 인공 지능 기법으로 인정받고 있다. 때문에 딥러닝은 새로운 개념이 아니며, 1965년 이후로 연구자들에 의해 지속적으로 연구되어져 왔다. 최근에 오버피팅(Over-fitting) 문제가 완화되고 하드웨어의 발전으로 인한 학습 시간의 문제가 어느 정도 해결됨에 따라, 딥러닝이 다시 주목받고 있다. 또한 많은 양의 데이터가 생산됨에 따라 학습 데이터의 양을 더 많이 확보할 수 있기 때문에 더 복잡한 개념과 표현을 쉽게 학습할 수 있게 되었다.

대표적인 딥러닝 알고리즘에는 Auto-encoder[9], Restricted Boltzmann Machine(RBM)[10], Convolutional Neural Network(CNN) 및 RNN[11] 등이다. Auto-encoder 및 RBM은 초기에 제안된 모델이고, 자가 학습이라는 장점이 있지만 성능은 제한적이다. 현재 대부분의 딥러닝 시스템들은 CNN 또는 RNN을 기반으로 한다.

CNN은 이미지 신호 처리 및 컴퓨터 비전 분야에서 없어서는 안 될 알고리즘이 되었으며[12], RNN은 순차적인 의미 정보를 처리하는데 적합하므로 음성 신호 처리 및 음성 인식에서 우수한 성능을 보인다. 최근에는 이러한 초기 알고리즘을 보완하거나 개선하는 많은 좋은 알고리즘들이 제시되고 있다.

우리는 이 연구에서 주로 사용될 RNN의 최신 연구 경향을 더 검토했다. RNN은 출력값의 일부가 입력 값에 다시 포함되는 연속성 개념의 신경 회로망을 의미한다. 이러한 종류의 신경망은 시계열 데이터를 처리하기 위해 제안되었지만, 장기간에 걸쳐 발생되는 패턴을 인식하지 못하는 문제점이 발견되었다. 이 문제를 해결하기 위해서, Long Short-Term Memory(LSTM)[13]과 Gated Recurrent Unit(GRU)[14] 등의 알고리즘이 제시되었다. LSTM은 가중치 뿐 아니라 메모리에 대한 추가 정보를 셀 상태에 저장하고 시계열 패턴의 길이도 조정한다. GRU는 LSTM의 출력 게이트를 제거하기 때문에 성능은 동일하게 유지하면서 보다 간단해졌다. RNN은 주로 음성 신호 처리[15] 및 문자열 처리[16]에 사용된다.

## Ⅲ. 연구 설계

### 1. 변수 선정

2.2장에서 기술한 것과 같이, 자연 재난에 따른 농산물 가격 예측을 위해 활용된 변수들은 기상(또는 날씨), 과거 가격 변화, 생산량/재배면적, 공급량, 유가, 농가소득 등이었다. 기상 정보는 자연 재해의 발생 여부를 알

아내기 위한 정보이며, 과거 가격 변화는 미래 예측을 위해 필수적으로 필요한 변수이다. 생산량/재배면적/공급량은 가격 형성에 영향을 미치는 주요 요소이며, 유가는 농기계 운영에 필요한 요소라서 농사 비용에 영향을 주는 변수라 할 수 있다. 농가소득은 도매가격 예측을 위해 사용된 변수인데, 농민들이 1차 유통업자들과 거래할 때 거래가격에 영향을 미치는 요소라서 도매가격 예측에 포함된 것으로 추측된다. 따라서, 농가소득은 소비자 가격에 미치는 영향은 적을 것으로 판단되고, 특정 지역 및 대상 농산물을 재배하는 농가소득에 대한 데이터를 구하기 어려웠기 때문에 제외하였다.

위와 같은 선행 연구들에서 사용된 변수들을 고려하여 본 연구에서는 농산물 소비자 가격 예측을 위해 [표 2]와 같은 변수들을 사용하였다.

**표 2. 농산물 가격 예측을 위한 변수**

| 구분 | 변수 |
|---|---|
| 기상 변수 | 기온, 강수량, 습도, 풍량, 적설량<br>※ 각 변수에 대해 최저값, 최고값, 평균값, 중간값 사용 |
| 기타 변수 | 경유, 물가 상승률, 전년 수확량, 전년 수입량,<br>전년 재배면적 |

기상 변수는 기온, 강수량, 습도, 풍량, 적설량으로 구성되며, 각각에 대해서 최저값, 최고값, 평균값, 중간값을 가진다. 경유 가격, 물가 상승률, 전년도 수확량, 전년도 재배면적, 수입량은 부가적인 변수의 성격을 가진다. 농산물 가격의 변화는 기상 정보만으로 설명할 수 없는 부분이 있기 때문에, 기상 정보를 보완할 수 있는 데이터로서 위와 같은 변수들도 학습에 고려할 필요가 있었다.

### 2. 알고리즘 선정

농산물의 가격은 일반적으로 계절을 기준으로 심한 변화를 보인다. 계절을 고려한 예측이 이루어져야 하지만, 한 계절은 너무 많은 기간을 가지기 때문에 학습의 단위로 적합하지 않다. 따라서, 계절보다는 짧지만 가격의 변화를 어느 정도 반영할 수 있는 한 달이란 기간을 기준으로 학습 데이터를 준비하였다. 즉, 과거 3주 21일간의 데이터로 학습 모델을 만들어서 이후 1주 7일간의 농산물 가격을 예측하는 태스크로 정의하였다. 학습 데이터 구축은 4장에서 자세히 설명할 예정이다.

농산물을 포함한 대부분의 판매 제품의 가격은 기본적으로 과거의 가격 변화와 무관하지 않다. 특별한 경우를 제외하고는 과거의 가격 변화에서 심하게 벗어나지 않는다. 따라서, 미래의 농산물 가격 예측을 위해서도 과거의 가격을 반영하는 것이 중요하기 때문에 과거의 정보를 기억해서 미래의 입력 정보로 넘겨주는 RNN 계열의 딥러닝 알고리즘이 적합하다고 판단된다.

본 연구에서는 딥러닝 알고리즘 중 시계열 분석에 적합한 RNN 계열의 LSTM 알고리즘을 사용하였다. LSTM 알고리즘은 RNN의 단점을 보완한 알고리즘이다. RNN은 시간 순서로 받아들인 입력데이터를 학습할 때 은닉층(hidden layer)에 기억 기능이 있어서 각각의 상태를 저장했다가 학습에 활용하는 신경망이다. RNN에서 과거 데이터의 기억 기능을 보강한 알고리즘이 LSTM이다[7]. LSTM에는 과거 기억을 저장하고 다음 셀로 흘려 보낼 기억의 양을 조절하는 Layer들이 있는데, Forget gate, Input gate, Update gate, Output gate들이다. [그림 1]은 본 연구에 적용된 LSTM의 아키텍처이며, 각 gate의 역할은 (1)~(4)에서 표현되고 있다. 알고리즘의 파라미터에 관련 부분은 Chapter 5에서 설명할 예정이다.

## Ⅳ. 학습 데이터 구축

### 1. 학습 데이터 수집 및 전처리

기상 데이터는 기상청[1]으로부터, 농산물 가격 데이터는 농산물 유통정보서비스(KAMIS)[2], 유가 데이터는 오피넷(Opinet)[3]으로부터 각각 수집하였다. 물가 상승률, 전년도 수확량, 재배면적, 전년도 수입량 데이터는 통계청 국가통계포털(KOSIS)[4]에서 데이터를 개방·공유하고 있다. [표 3]에서 학습에 사용된 데이터의 출처를 정리하였다.

1 https://data.kma.go.kr/cmmn/main.do
2 https://www.kamis.or.kr/customer/main/main.do
3 http://www.opinet.co.kr/user/main/mainView.do
4 http://kosis.kr/index/index.jsp

**표 3. 학습 데이터 출처**

| 구분 | 출처 |
|---|---|
| 기상 데이터 | 기상청 |
| 농산물 가격 데이터 | 농산물 유통정보서비스 |
| 유가 | 오피넷 |
| 물가 상승률, 전년도 수확량, 재배면적, 전년도 수입량 | 국가통계포털 |

**표 4. 학습 데이터**

| 변수 | 데이터 |
|---|---|
| 가격(1) | 농산물가격(1) |
| 기상(25) | 평균기온(광주/대구/대전/부산/서울)(5)<br>평균강수량(광주/대구/대전/부산/서울)(5)<br>평균습도(광주/대구/대전/부산/서울)(5)<br>평균풍량(광주/대구/대전/부산/서울)(5)<br>평균적설량(광주/대구/대전/부산/서울)(5) |
| 유가(16) | 서울/부산/대구/인천/광주/대전/울산/경기/강원/충북/충남/전북/전남/경북/경남/제주 |
| 물가 상승률(33) | 총지수, 식료품/비주류음료, 식료품, 빵및곡물, 육류, 어류및수산, 우유/치즈및계란, 식용유지, 과일, 채소및해조, 배추, 무, 파, 양파, 과자/빙과류 및 당류, 기타식료품, 비주류음료, 주류및 담배, 의료및신발, 주택/수도/전기및연류, 전기/가스및기타연료, 가정용품및가사서비스, 보건, 교통, 운송장비, 개인운송장비운영, 운송서비스, 통신, 오락및문화, 교육, 음식및숙박, 음식서비스, 기타상품및서비스 |
| 시도별 전년수확량(16) | 서울/부산/대구/인천/광주/대전/울산/경기/강원/충북/충남/전북/전남/경북/경남/제주 |
| 시도별 전년 재배면적(16) | 서울/부산/대구/인천/광주/대전/울산/경기/강원/충북/충남/전북/전남/경북/경남/제주 |
| 전년 수입량(1) | 전년 수입량 |

각각의 소스로부터 공개 가능한 데이터들을 수집하였고, [표 5]에서 정리하였다. 수집한 농산물 가격 데이터, 기상 데이터, 부가 데이터들은 기간이 달랐다. 예를 들면, 기상 데이터는 1999년 1월부터 시작되고, 물가 상승률은 1997년부터 존재한다. 때문에, 효과적인 학습을 위해서는 공통으로 존재하는 기간의 데이터만 활용하는 것이 좋다. 모든 데이터들이 2000년 1월 1일부터 2016년 12월 31일 사이에 데이터 값이 존재하기 때문에, 이 기간을 기준으로 학습 데이터를 준비하였다. 데이터마다 이 기간 중 비어 있는 값들이 있는지 확인하고, 값들을 채우는 작업을 수행하였다. 예를 들어, 주말과 공휴일에는 데이터 값이 없는 경우가 많아서 이 경우에는 앞날과 뒷날의 데이터 값의 평균으로 채워주었다. 전년 생산량과 같이 연 단위로 기록된 데이터들은 동일한 해의 날들에는 동일한 데이터 값으로 채워주었다. 학습 데이터는 [표 6]과 같이 정리된다.

변수와 실제 수집 및 전처리 된 데이터를 고려했을 때 세부적으로 사용된 변수는 전체 108개이며 [표 4]와 같이 정리했다.

## 2. 학습 데이터 포맷

[그림 1]은 본 연구에서 사용된 LSTM 네트워크의 구조를 보여준다. LSTM에는 Forget gate, Input gate, Update gate, Output gate 등 과거 메모리를 저장하고 다음 셀로 보낼 메모리 양을 제어하는 레이어가 있다. 그림에서 (1), (2), (3), (4)는 각각의 기능을 설명하고 있다. $Ct$는 셀 상태를 나타내며, 이전 상태로부터 전송된 장기 정보를 다음 단계로 전송한다. Forget gate(1)는 셀 상태에서 버리고 유지할 정보를 결정한다. $f_t$ 함수는 이전 셀 상태($h_{t-1}$)와 현재 셀 입력 값 $x_t$의 조합을 $\sigma$ 한 값이며, '0'과 '1' 사이의 값을 갖는다. '0'은 정보를 완전히 제거하는 것을 의미하고 '1'은 정보를 온전히 유지하는 것을 의미한다. Input gate(2)는 새로운 정보가 셀 상태에 저장될지를 결정하는 계층이다. 이것은 $\sigma$ 또는 *tanh*를 통해 $h_{t-1}$과 $x_t$를 결합하여 새로운 입력 값을 만든다. 이 입력 값은 다음 계층에서 수행 할 업데이트 대상이 된다. Update gate(3)에서, 셀 상태 값은 Input gate 값을 Forget gate 값과 결합하여 업데이트 된다. 마지막으로 출력 값은 업데이트 된 셀 상태 값의 *tanh* 값을 Output gate(4)에 반영하여 결정된다. 이 연구에서는 구글의 텐서플로우 플랫폼에서 제공하는 BasicLSTMCell API를 가져와 LSTM 네트워크를 구현하는 데 사용했다.

시계열 데이터 예측에 유용한 LSTM 네트워크를 사용하기 위해, 학습 데이터의 형식은 $M \times N$ 형식의 매트릭스로 표현되어야 한다. 본 연구에서는 108개의 입력 데이터에 대해 21일치 데이터를 사용하여 $108 \times 21$ 매트릭스를 학습 데이터 포맷으로 사용했다. 출력 결과가 7일 동안의 농산물 가격이므로 출력 포맷은 $1 \times 7$ 매트릭스가 된다[그림 2].

**표 5. 초기 수집 데이터**

| 변수 | 세부 변수 | 기록 단위 | 지역 | 기간 | 비고 |
|---|---|---|---|---|---|
| 기상 변수 | 기온 | 일간 | 전국 지역별 | 1999.01~현재 | – |
| | 강수량 | 일간 | 전국 지역별 | 1999.01~현재 | – |
| | 습도 | 일간 | 전국 지역별 | 1999.01~현재 | – |
| | 풍량 | 일간 | 전국 지역별 | 1999.01~현재 | – |
| | 적설량 | 일간 | 전국 지역별 | 1999.01~현재 | – |
| 농산물 가격 변수 | 쌀 | 일간 | 광주, 대구, 대전, 부산, 서울 | 1999.01~현재 | 상품, 중품 |
| | 시금치 | 일간 | 광주, 대구, 대전, 부산, 서울 | 1999.01~현재 | 상품, 중품 |
| | 애호박 | 일간 | 광주, 대구, 대전, 부산, 서울 | 1999.01~현재 | 상품, 중품 |
| | 대파 | 일간 | 광주, 대구, 대전, 부산, 서울 | 1999.01~현재 | 상품, 중품 |
| | 양파 | 일간 | 광주, 대구, 대전, 부산, 서울 | 1999.01~현재 | 1kg/20kg, 햇양파/양파, 상품/중품 |
| 부가 변수 | 경유 | 일간 | 16개 시 · 도 | 1997.01~현재 | – |
| | 전년 물가 상승률 (농산물 관련 34개 품목) | 연간 | 16개 시 · 도 | 1997.01~현재 | – |
| | 전년 수확량 | 연간 | 16개 시 · 도 | 2001.01~현재 | 10a 당 수확량, 생산량(톤) |
| | 전년 수입량 | 연간 | 전국 | 2000.01~현재 | – |
| | 전년 재배면적 | 연간 | 16개 시 · 도 | 1997.01~현재 | 논,밭, 합계 |

**표 6. 전처리 후 데이터**

| 변수 | 세부 변수 | 기록 단위 | 지역 | 기간 | 전처리 내용 |
|---|---|---|---|---|---|
| 기상 변수 | 기온 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 예측 대상 지역만 선별 |
| | 강수량 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 예측 대상 지역만 선별 |
| | 습도 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 예측 대상 지역만 선별 |
| | 풍량 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 예측 대상 지역만 선별 |
| | 적설량 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 예측 대상 지역만 선별 |
| 농산물 가격 변수 | 쌀 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 중품 기준 |
| | 시금치 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 중품 기준 |
| | 애호박 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 중품 기준 |
| | 대파 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 중품 기준 |
| | 양파 | 일간 | 광주, 대구, 대전, 부산, 서울 | 2000.01.01~2016.12.31 | 1kg/양파/중품 기준 |
| 부가 변수 | 경유 | 일간 | 16개 시 · 도 | 2000.01.01~2016.12.31 | – |
| | 전년 물가 상승률 (농산물 관련 34개 품목) | 일간 | 16개 시 · 도 | 2000.01.01~2016.12.31 | 연간을 일간으로 적용 |
| | 전년 수확량 | 일간 | 16개 시 · 도 | 2000.01.01~2016.12.31 | 연간을 일간으로 적용 |
| | 전년 수입량 | 일간 | 전국 | 2000.01.01~2016.12.31 | 연간을 일간으로 적용 |
| | 전년 재배면적 | 일간 | 16개 시 · 도 | 2000.01.01~2016.12.31 | 연간을 일간으로 적용 |

※ 모든 변수 데이터에 대해서 '기간'을 2000년 1월 1일부터 2016년 12월 31일로 일괄적으로 적용

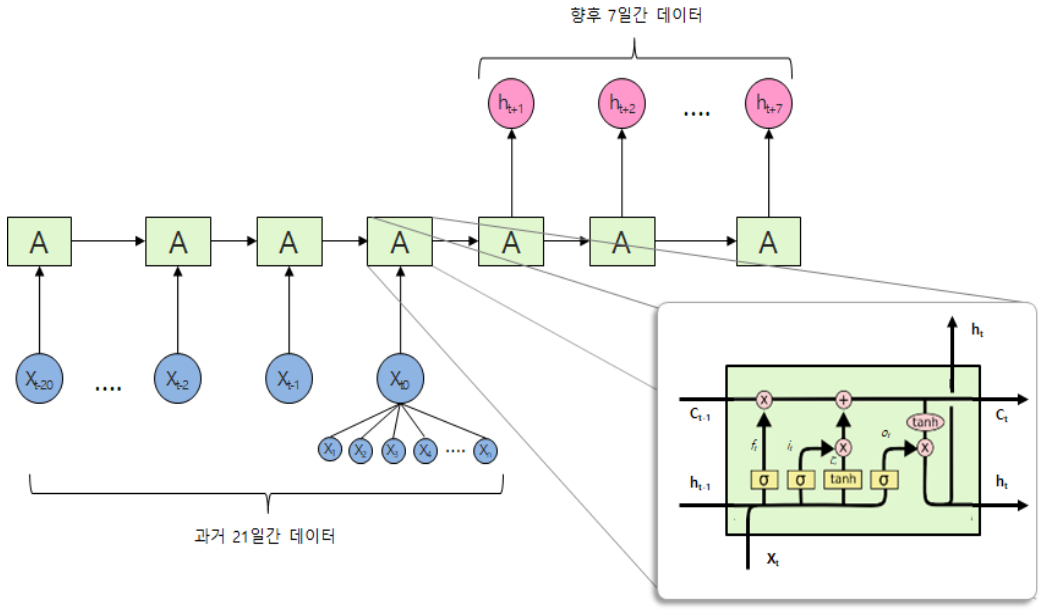


그림 1. 활용된 LSTM 네트워크 구조

$$\text{Forget gate}: \quad f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{1}$$

$$\text{Input gate}: \quad i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right),\ \widetilde{C}_t = tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{2}$$

$$\text{Update gate}: \quad C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{3}$$

$$\text{Output gate}: \quad o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right),\ h_t = o_t * \tanh(C_t) \tag{4}$$

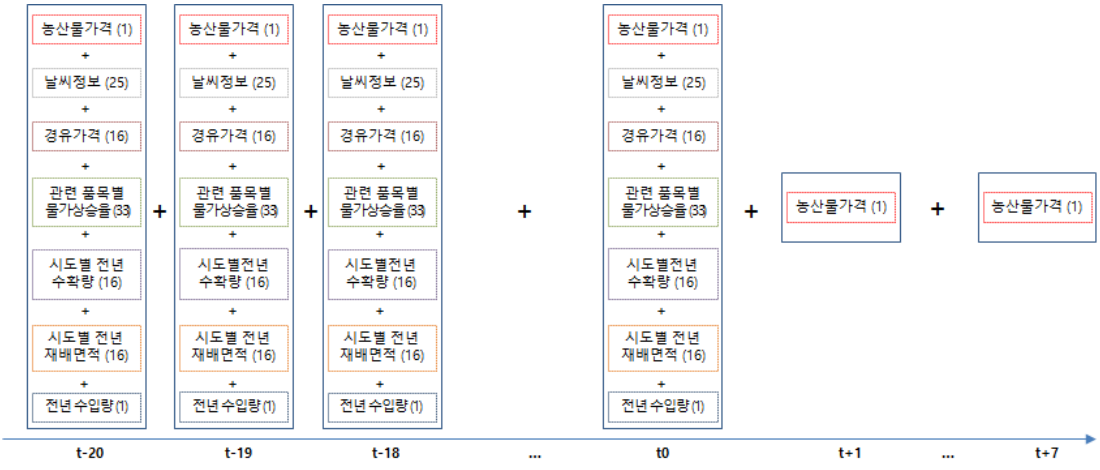


그림 2. 학습 데이터 구조

## V. 실험 및 결과

2000년 1월 1일부터 2016년 12월 31일까지의 기간에서 21일간 데이터를 하나의 인스턴스로 했을 때 학습 데이터의 전체 인스턴스 수는 6,182개이다. 학습 및 테스트를 위해 전체 인스턴스의 80%가 학습 모델을 만드는 데 사용되었으며, 나머지 20%로 학습 모델의 정확성을 테스트했다. [표 7]은 학습 및 테스트 데이터의 건수를 나타낸다.

**표 7. 학습/테스트 데이터셋 건수**

| 구분 | 건수 |
|---|---|
| training | 4,945 |
| testing | 1,237 |
| 합계 | 6,182 |

입력 데이터의 값이 가격, 비율, 온도, 면적 등 종류가 다양하기 때문에 정확한 학습을 위해 데이터 정규화 작업을 해 주었다. (5)와 같은 일반적인 정규화 식을 통해 모든 학습 데이터에 대해서 0~1 사이의 값으로 변환해서 정규화 된 값으로 학습과 테스트를 진행하였다.

$$f(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (5)$$

좋은 예측 모델을 얻으려면 LSTM 네트워크의 파라미터들의 값 설정이 중요하다. Learning rate, number of iteration, mini-batch size 등과 같은 parameter들을 네트워크 내에서의 파라미터와 구별되도록 hyper-parameter라 부른다.

Learning rate은 예측 값과 실제 값의 차이를 나타내는 cost function이 그리는 그래프 상에서 한번에 값이 이동하는 거리를 나타낸다. 한 번에 너무 많은 거리를 이동하면 cost function의 결과를 최소화하는 매개변수 W 값을 지나칠 수 있으며, 너무 적은 거리를 이동하면 찾는데 학습에 너무 많은 시간이 소요되어 비효율적이다. 많은 기존의 연구들에서, 학습 속도는 일반적으로 0.001로 설정한다.

딥러닝에서는 한번에 활용하는 학습 데이터의 크기를 결정하는데 있어서, 기존의 기계학습에서 사용하는 full batch 방식과는 다르게 mini-batch 방식을 활용한다. Full batch 방식은 한번의 학습에 모든 데이터를 활용해서 학습 모델을 만드는 방법인 반면, mini-batch 방식은 전체 데이터를 학습에 활용하되 한번의 과정(iteration) 내에서 전체 데이터를 서브 데이터셋으로 나누어 여러 번 학습한다. 학습에 사용되는 데이터가 분석 대상이 되는 모집단 전체가 될 수 없기 때문에 랜덤하게 서브 데이터셋으로 나누어 여러번 학습 해 주는 것이 모집단 전체를 더 잘 대표할 수 있다는 장점이 있다. 절대적인 mini-batch 크기는 존재하지 않지만, 보통 총 학습 데이터의 약 1 ~ 2 % 정도를 mini-batch 크기로 설정한다. 본 연구에서는 [표 7]에서 제시된 학습셋 4,945건에 대해서 mini-batch의 크기를 10으로 하고, 한번의 epoch에서 495회의 학습 과정(iteration)을 거치도록 설계하였다.

학습 시 학습 데이터에만 최고의 성능을 보이는 학습 모델이 만들어지는 문제점을 over-fitting이라 한다. 즉, 학습 데이터에 대해서는 에러가 작지만, 일반적인 데이터에 대해서는 에러가 커지는 문제이다. Over-fitting 문제를 해결하기 위해 여러 방법들이 제시되고 있는데, 본 연구에서는 비교적 좋은 해결 방법으로 알려진 drop-out 기법을 사용했고, 그 값을 0.01로 주었다.

위에서 언급했던 hyper-parameters 값들을 [표 8]에서 정리하였다.

**표 8. hyper-parameters 값**

| Hyper-parameters | 값 |
|---|---|
| Learning rate | 0.001 |
| Mini-batch size | 10 |
| Number of iteration | 500 |
| Drop-out value | 0.01 |

정확도가 높은 예측 모델을 얻기 위해서는 hyper-parameters 값을 결정하는 것과 더불어, cost

function을 최소화하는 optimizer를 선정하는 것도 중요하다. 딥러닝 연구의 초기에는 gradient descent optimizer를 주로 사용하였는데, 이후 더 좋은 성능을 보이는 adam optimizer가 제시되고 그후로도 더 개선된 optimizer들이 많이 연구가 되었다. 본 연구에서는 일반적으로 많이 사용되는 adam optimizer를 사용하였다.

Epoch는 학습의 반복을 의미하며 가장 뛰어난 예측 모델을 만든 시점에서 중지해야 한다. 컴퓨팅 리소스가 제한적이고 학습 시간에 대한 제약으로 인해 조기에 중지 지점을 찾는 것이 중요하다. 이를 위해, epoch 수와 loss function의 결과 사이의 상관 관계 그래프를 그려서 epoch의 최적값을 결정할 수 있다. 본 연구에서는 실험을 통해 epoch 수가 400~500회 사이에서 가장 좋은 loss function의 결과를 보였고, 500회 이상에서는 loss 값의 변화가 거의 없음을 확인했다.

모델에 의한 예측 가격과 실제 가격의 차이를 통해 모델의 성능을 가늠해 볼 수 있다. 두 값의 오차를 측정할 수 있는 가장 보편적인 평가 도구는 Mean Squared Error(MSE)와 MSE의 제곱근인 Root mean squared error(RMSE)를 사용하여 왔다. 예측이라는 것은 음의 결과도 나올 수 있으므로, MSE보다는 RMSE가 더 적합하여 본 연구에서는 오차 측정 도구로 RMSE를 사용하였다(6).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2} \qquad (6)$$

위와 같은 학습 데이터와 실험 환경으로 농산물 가격 예측 모델의 정확도 실험 결과는 [표 9]와 같다. 예측 대상은 대구, 광주, 서울, 대전, 부산 등 5개 도시별로 5개 농산물들의 7일간의 가격이다.

농산물별로는 대파에 대한 정확도가 가장 좋고, 애호박은 가장 낮은 예측 정확도를 보이고 있다. 예측 정확도를 기준으로 농산물을 분류해 보면, 대파, 양파, 쌀이 비교적 높은 정확도를 보이고 있고, 그 다음은 시금치, 마지막으로 애호박으로 분류할 수 있다. 시금치나 애호박이 대파, 양파, 쌀에 비해 정확도가 낮은 원인을 대중성에서 찾아볼 수 있을 것으로 추측한다. 일반적으로 대파, 양파, 쌀은 애호박과 시금치에 비해 요리의 재료로서 활용성이 적다고 볼 수 있다. 이런 원인으로 생산 및 거래량이 상대적으로 적어서 가격 변화가 심할 수 있다고 판단된다. 실제로 1, 3, 5, 7일의 가격 예측 결과인 [그림 3-6]의 농산물별 가격 변동성을 보더라도 애호박과 시금치의 가격 변화 파동이 심한 것을 볼 수 있다.

[표 9]는 구체적인 정확도를 수치로 나타내기 때문에 이 수치들이 어느 정도 정확한지를 판단하기가 쉽지 않다. [그림 3-6]은 실 데이터와 예측 데이터를 비교함으로서 예측 정확도를 시각적으로 표현하였고, 모델에 의한 예측 가격이 비교적 실제 가격을 잘 추종하는 것으로 볼 수 있다.

**표 9. 도시별/농산물별 가격 예측 정확도**

| 지역 | 대파 | 양파 | 애호박 |
|---|---|---|---|
| 대구 | 0.047002 | 0.059741 | 0.098147 |
| 광주 | 0.053291 | 0.050698 | 0.090949 |
| 서울 | 0.054660 | 0.072942 | 0.077357 |
| 대전 | 0.057111 | 0.053470 | 0.121505 |
| 부산 | 0.051233 | 0.082001 | 0.096845 |
| 합계 | 0.052659 | 0.06377 | 0.096961 |
| 지역 | 쌀 | 시금치 | *전체 합계* |
| 대구 | 0.056684 | 0.085620 | *0.069439* |
| 광주 | 0.059716 | 0.071724 | *0.065276* |
| 서울 | 0.052059 | 0.070472 | *0.065498* |
| 대전 | 0.059265 | 0.074336 | *0.073137* |
| 부산 | 0.060045 | 0.076212 | *0.073267* |
| 합계 | 0.057554 | 0.075673 | *0.069323* |

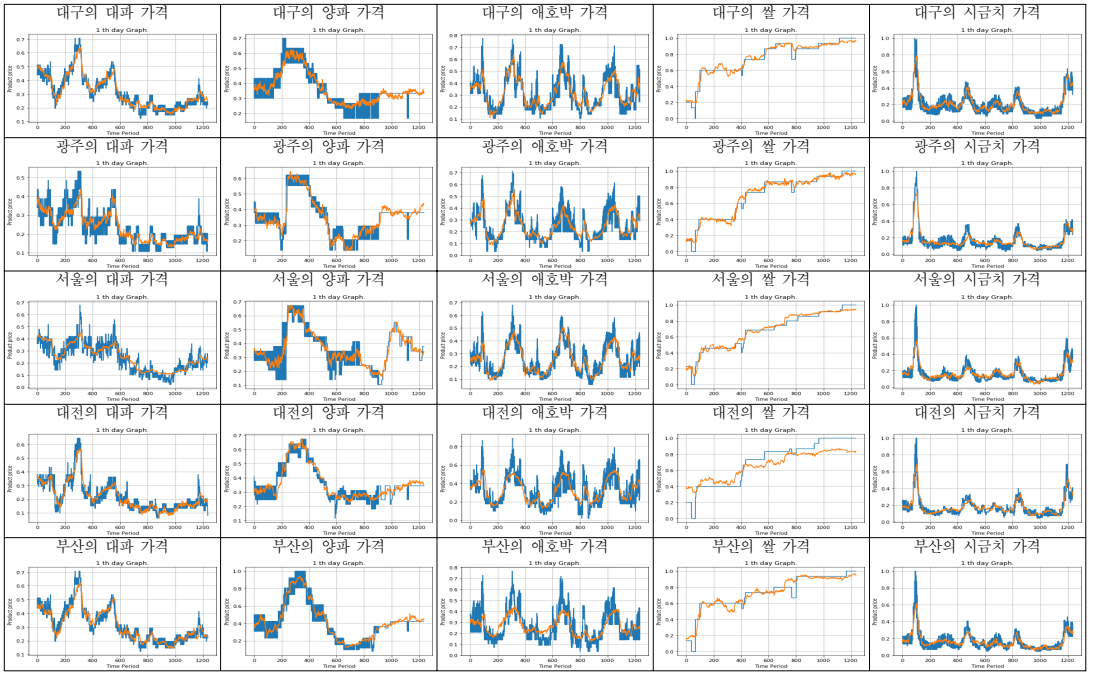


그림 3. (1일차) 도시별/농산물별 예측 정확도 그래프

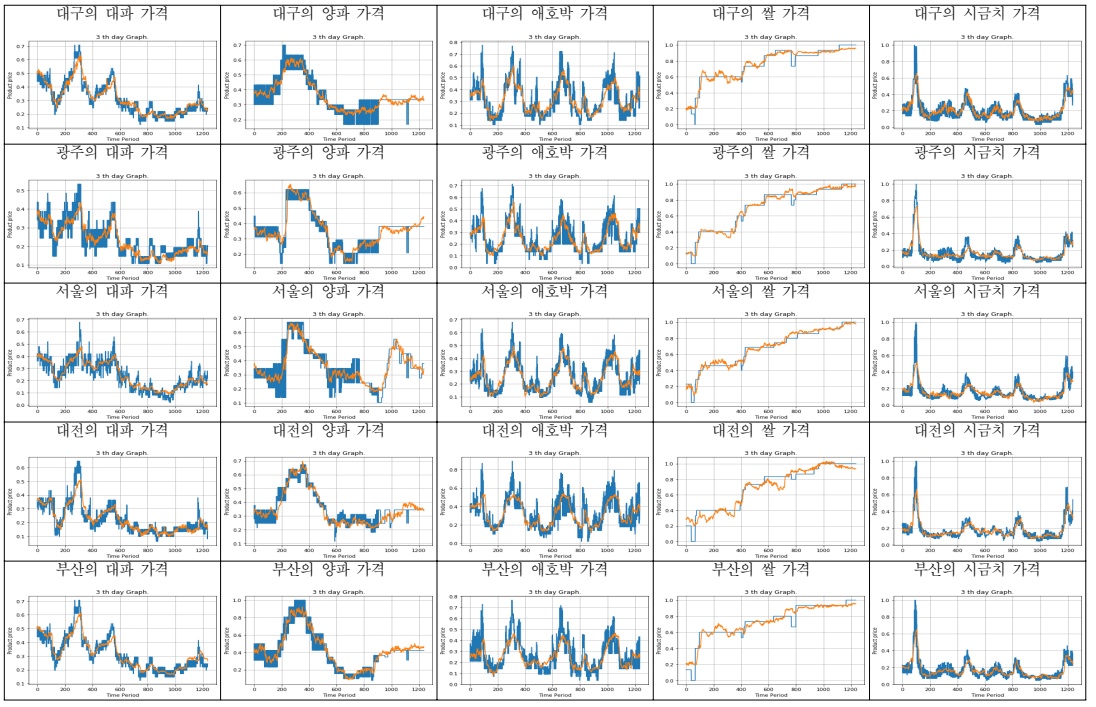


그림 4. (3일차) 도시별/농산물별 예측 정확도 그래프

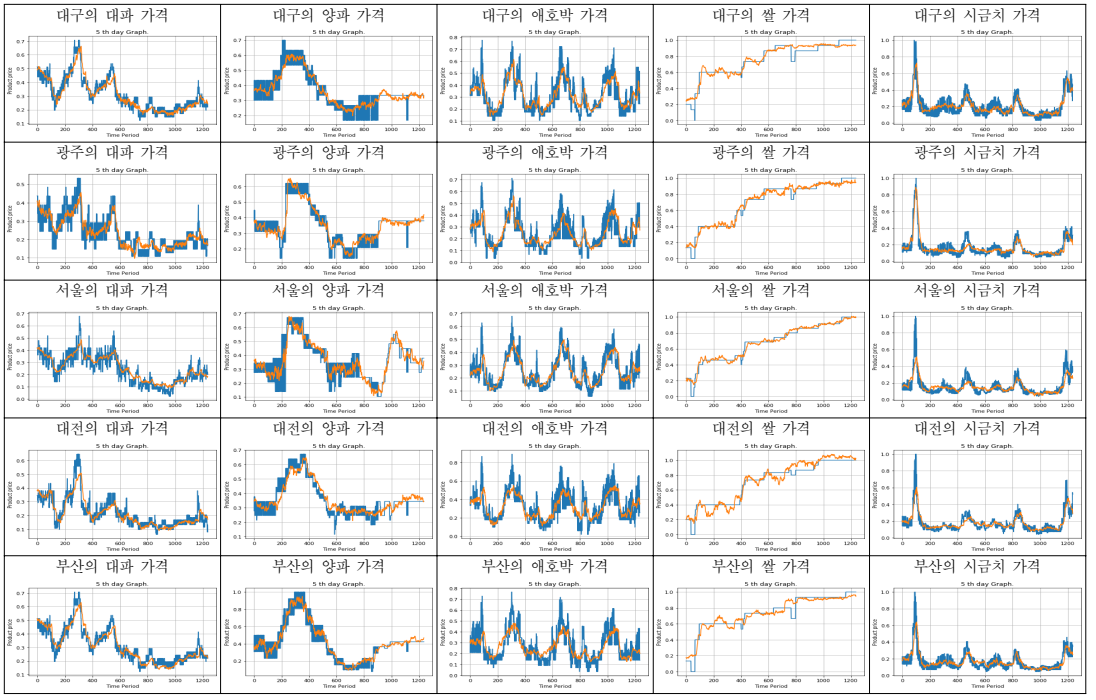


그림 5. (5일차) 도시별/농산물별 예측 정확도 그래프

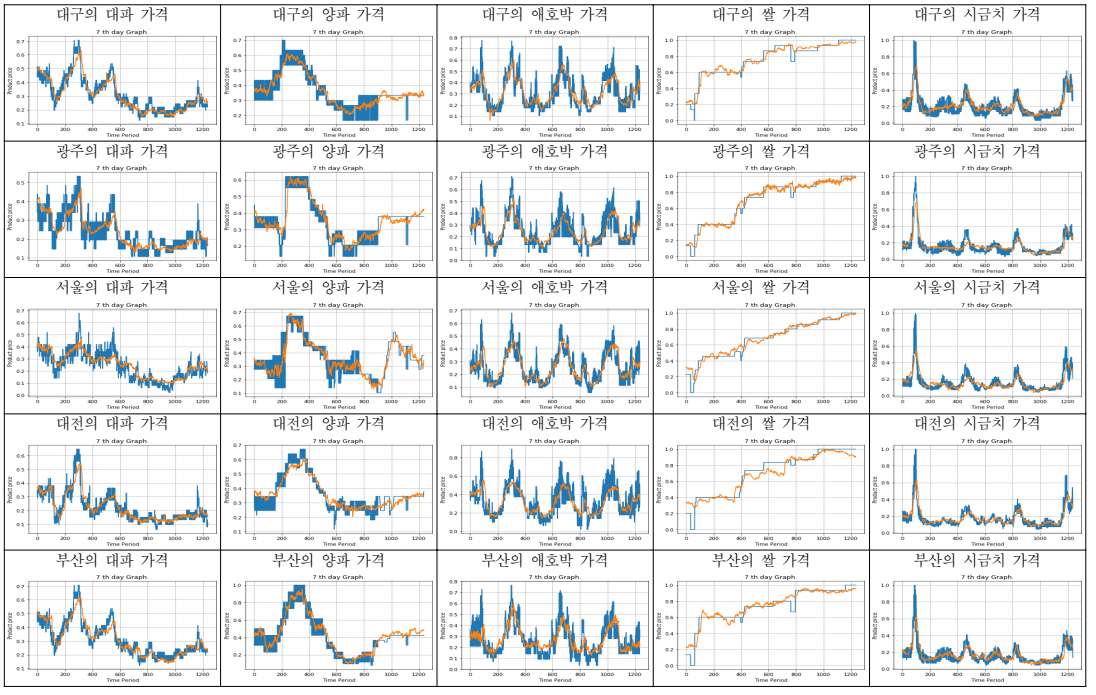


그림 6. (7일차) 도시별/농산물별 예측 정확도 그래프

## Ⅵ. 결론

본 연구의 목적은 자연 재해로부터 받을 수 있는 여러 영향 중 농산물 가격 변화에 대한 효과적인 대응 방안을 수립하기 위한 딥러닝 기반의 농산물 가격 예측모델을 제안하는 것이다.

농산물 가격 예측 모델의 개발을 위해, 기상 데이터, 농산물을 포함한 관련 물품들의 가격 인상률, 유가, 농산물별 전년도 수확량, 농산물별 전년도 재배 면적 등 농산물 가격 변화에 영향을 미치는 변수들에 대한 17년치 데이터를 수집하여 학습에 활용하였다. Learning rate, Mini-batch size, Number of iteration, Drop-out value 등 Hyper-parameters의 값을 설정하고 Adam Optimizer를 사용해서 LSTM 네트워크를 적용한 결과, 모델의 정확도는 백분율 기준으로 약 93%이며, 비교적 농업 가격 예측을 잘 할 수 있는 것으로 판단된다. 이 결과는 이전의 시계열 데이터 예측 모델이나 회귀분석 모델들이 보여주는 90% 수준의 예측력보다 높은 수준이다.

본 연구에서 제안하는 농산물 가격에 대한 딥러닝 기반 예측 모델은 농산물 가격 동향 지수의 형태로 제공되어 농산물 수급을 위한 정부 정책 수립에 사용될 수 있고, 일반 소비자들에게도 유용한 정보로 활용될 수 있을 것으로 기대된다.

향후에는 실제 자연재해가 발생한 기간이나 그 이후의 시점에 초점을 맞추어 이 기간 동안의 예측 정확도가 어느 정도인지 분석하는 연구가 진행될 예정이다.

## 참 고 문 헌


[1] M. Fafchamps and B. Minten, “Impact of SMS-Based Agricultural Information on Indian Farmers,” In the World Bank Economic Review, Vol.26, Iss.3, pp.383-414, 2012.

[2] A. M. Rather, A. Agarwala, and V. N. Sastryb, “Recurrent Neural Network and a Hybrid Model for Prediction of Stock Returns,” In Expert Systems with Applications, Vol.42, pp.3234-3241, 2015.

[3] H. Chiroma, S. Abdulkareem, and T. Herawan, “Evolutionary Neural Network Model for West Texas Intermediate Crude Oil Price Prediction,” In Applied Energy, Vol.142, pp.266-273, 2015.

[4] J. H. Zhang, F. T. Kong, J. Z. Wu, M. S. Zhu, K. Xu, and J. J. Liu, “Tomato Prices Time Series Prediction Model Based on Wavelet Neural Network,” In Applied Mechanics & Materials, Vol.644-650, pp.2636-2640, 2014.

[5] C. Wang, A. Zhao, and Y. Zhao, “Design and Implementation of Agricultural Product Prices Short-Term Forecasting System,” Proceedings of 2013 World Agricultural Outlook Conference, pp.15-27, 2013.

[6] 임지연, *시계열 모형을 이용한 중기 농산물가격 예측 분석 : 백다다기오이와 애호박을 중심으로*, 중앙대학교, 석사학위논문, 2015.

[7] 남국현, “양파 출하시기 도매가격 예측모형 연구,” 농촌지도와 개발, 제22권, 제4호, pp.423-434, 2015.

[8] 배경태, *인공신경망 기법을 이용한 최적의 농산물 가격 예측모델 개발*, 숭실대학교, 석사학위논문, 2017.

[9] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion,” In Journal of Machine Learning Research, Vol.11, pp.3371-3408, 2010.

[10] G. E. Hinton, S. Osindero, and Y. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” In Neural computation, Vol.18, No.7, pp.1527-1554, 2006.

[11] T. Mikolov, S. Kombrink, and L. Burget, “Extensions of Recurrent Neural Network Language Model,” Proceedings of 2011 IEEE

International Conference on Acoustics, Speech and Signal Processing, pp.5528-5531, 2011.

[12] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.39, No.4, 2017.

[13] F. A. Gers, J. A. Schmidhuber, and F. A. Cummins, "Learning to Forget: Continual Prediction with LSTM," In Neural Computation, Vol.12, No.10, pp.2451-2471, 2000.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," arXiv:1412.3555, 2014.

[15] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid Speech Recognition with Deep Bidirectional LSTM," In Automatic Speech Recognition and Understanding, pp.273-278, 2013.

[16] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask Me Anything: Dynamic Memory Networks for Natural Language Processing," 2015.
http://arxiv.org/abs/1506.07285

## 저 자 소 개

**신 성 호(Sungho Shin)** 정회원

- 2002년 8월 : 경북대학교 경영정보학 전공(이학석사)
- 2014년 2월 : 한국과학기술원 지식공학 박사 수료
- 2002년 9월 ~ 현재 : 한국과학기술정보연구원 연구데이터플랫폼센터 선임연구원

<관심분야> : 딥러닝, 인공지능, 정보추출, 연구데이터

**이 미 경(Mikyoung Lee)** 정회원

- 2002년 2월 : 경북대학교 컴퓨터공학과(공학석사)
- 2002년 2월 ~ 2005년 6월 : 한국전자통신연구원 지식및추론연구팀 연구원
- 2005년 6월 ~ 현재 : 한국과학기술정보연구원 연구데이터플랫폼센터 선임연구원

<관심분야> : 빅데이터, 딥러닝, HCI, 연구데이터플랫폼

**송 사 광(Sa-kwang Song)** 정회원

- 2011년 2월 : 한국과학기술원 전산학과(공학박사)
- 2010년 12월 ~ 현재 : 한국과학기술정보연구원 연구데이터플랫폼센터 책임연구원
- 2014년 1월 ~ 현재 : 과학기술연합대학원대학교 빅데이터과학과 교수

<관심분야> : 딥러닝, 텍스트 마이닝, 자연어처리, 정보검색, 기계학습, 시맨틱 웹