

# 다중 선형 회귀 모델을 적용한 채소 가격 예측 시스템 : 기상 변인 중심으로

좌민혜\*, 연광제\*\*, 이재은\*\*\*, 김광용\*\*\*\*

## Vegetable Price Prediction System Applying Multiple Linear Regression Models : Focused on Meteorological Variables

Min-Hye Jwa\*, Kwang-Je Yeon\*\*, Jae-Eun Lee\*\*\*, Gwang-Yong Kim\*\*\*\*

### 요 약

채소 가격의 상승과 하락은 기상 여건에 의한 변동이 가장 크다. 특히 최근 48일째 이어진 장마철 집중호우와 태풍에 의한 채소 가격 상승은 가정의 장바구니 물가와 가공 및 외식업체의 경영상 위험요인이 된다. 따라서 본 논문은 기상 변인 중심의 채소 품목별 가격 예측 시스템을 제안한다. 예측 시스템은 사용자가 웹 사이트에 접속하여 기후 정보를 입력하면 예측한 채소 가격 정보를 사용자에게 제공한다. 제안하는 예측 시스템을 다중 선형회귀(Multiple Linear Regression) 모델을 적용한 딥러닝(Deep Learning) TensorFlow Library, Flask Web Server, MDBootstrap Design Framework, 기상청과 한국농수산식품유통공사 공공 데이터를 활용해 구현하고 실제 환경에서 적용 가능하다.

### ABSTRACT

The weather conditions are the most influencing factor of the rise and fall of vegetable prices. Recently, the rising vegetable prices due to 48 days of heavy rain and typhoons during the rainy season are risk factors for the domestic shopping cart prices and the management of processing restaurant businesses. Therefore, the thesis proposes a price prediction system for each vegetable based on weather variables. The prediction system provides the predicted vegetable price information to the user when the user accessing the website and inputs the weather information. The proposing prediction system can be implemented with the Deep Learning that has applied Multiple Linear Regression, TensorFlow Library, Flask Web Server, MDBootstrap Design Framework, and the public data from the Meteorological Agency and the Korea Agro-Fisheries & Food Trade Corporation and applied in the real environment.

키워드 : 예측 시스템, 다중 선형회귀, 딥러닝, 공공 데이터, 텐서플로우

key Words : Forecasting system, Multiple Linear Regression, Deep Learning, Public Data, TensorFlow

### I. 서 론

금년 한 해는 전례가 없는 코로나19 인한 경제 침체와 역대급 최장 장마, 태풍의 영향으로 채소 가격이 치솟아 소비자들이 느끼는 체감 물가는 훨씬 높았다. 한 신문 매체에 의하면 올해 배추의 경우, 생산량이 지난해보다 10%가량 줄면서 가격이 지난해

의 거의 배 수준으로 9월 기준 배추 소매가격은 상품 기준 포기당 1만310원으로 지난해 같은 달 5천 362원의 거의 배로 상승하여 가정과 기업 경영에 큰 부담을 안겨 주었다[1].

채소 가격의 주요 변동은 재배면적, 정부의 수급관리, 기상환경, 수출입 등 다양한 변화에 결정된다[2]. 그중 그림 1과 같이 기상변화에 의한 계절별 봄철(가

\* 송실대학교(lucia88@hanmail.net), \*\* (주)아이티노매즈(kjyeon@gmail.com), \*\*\* 널리지트브(leesunshin@soongsil.ac.kr),

\*\*\*\* (교신저자)송실대학교(gygim@ssu.ac.kr)

§ 논문번호 : 211303, 접수일자 : 2020년 10월 20일, 수정일자 : 2020년 11월 28일, 심사완료일자 : 2020년 12월 27일

몸, 고온), 여름과 가을철(집중호우, 태풍), 겨울철(폭설, 한파)에 가격 상승이 가장 컸다.

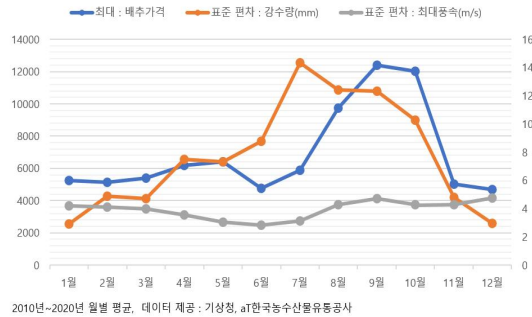


그림 1. 배수 최고가시 강수량과 풍속

가정과 음식점의 기본 식자재 채소 가격이 하락하면 생산 농가의 경영상 어려움이 발생하고 반대로 채소 가격이 상승하면 소비자의 가격 부담이 커지게 된다. 이러한 이유로 표준화된 데이터 기준의 가격 정보를 제공하는 기상 변인 중심의 채소 품목별 가격 예측 시스템을 제안한다. 생산자는 예측 시스템을 통해 경영 전략을 세우고 리스크를 최소화하며 소비자는 합리적인 구매를 결정할 수 있다. 또한 기존 가격 예측 연구들은 정해진 기간에 데이터를 학습 후, 해당 연도, 분기별 가격을 예측하였다. 반면 제안하는 가격 예측 시스템은 사용자가 웹 브라우저를 통해 채소 품목을 직접 선택하고 날씨 정보를 입력하는 사용자 중심의 서비스이다. 제안하는 시스템 설계를 위해 다중 선형회귀 모델을 적용한 딥러닝 알고리즘으로 학습하고 플라스크(Flask) 웹 서버, 텐서플로우(TensorFlow) 라이브러리, MDBootstrap 디자인 프레임워크, 기상청과 aT한국농수산식품유통공사 공공 데이터를 사용한다.

본 논문의 구성은 다음과 같다. 2장에서는 선형회귀와 딥러닝 선행연구를 소개한다. 3장에서는 제안하는 가격 예측 시스템을 기술하고 4장에서는 실험 결과를 기술한다.

## II. 관련연구

가격 변동은 일정한 시간의 흐름에서 독립 변수들의 변화에 가격이라는 종속 변수가 영향을 받는 것이다. 가격 예측에서는 회귀분석(Regression Analysis)과 시계열 분석(Time Series Analysis) 기법을 주로 활용한다. 회귀분석은 독립 변수와 종속 변수 사이의 상관관계를 나타내며, 독립 변수가 하나인 경우를 단

순 회귀분석, 여러 개인 경우를 다중 회귀분석이라 한다[3]. 시계열 분석은 특정 대상의 시간적 변동을 연속적으로 관측하여 얻는 자료에 근거하여 변동의 원인을 해명하고 미래를 예측하는 것을 의미한다[4]. 수학 및 통계 바탕을 두고 있는 전통적인 회귀분석이나 시계열 분석 연구 방법에서는 연구대상인 모집단 중 일부분의 샘플 데이터를 활용해 통계 도구들 또는 R 프로그래밍을 사용하여 수행한다. 이에 반해 딥러닝 기반의 회귀분석이나 시계열 분석 연구 방법에서는 인간의 사고방식과 유사한 방식으로 학습 모델을 통해 문제를 해결하고 예측한다.

### 1. 딥러닝 알고리즘

딥러닝(Deep Learning)은 인공지능 분야의 인공 신경망(ANN, Artificial Neural Network) 모형에서 발전한 형태로써 인공신경망의 내부계층이 여러 단계로 이루어진 그림 2와 같은 구조이다[5].

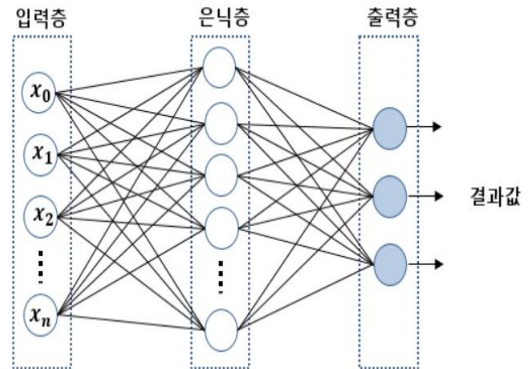


그림 2. ANN 구조

ANN의 구조는 데이터를 입력받는 입력층, 결과값을 출력하는 출력층, 최적의 Weight와 Bias를 찾아내는 은닉층으로 구성되어 있다. 은닉층의 개수와 노드의 개수 구성이 학습 모델이 된다. 이러한 모델을 ANN에서는 다층 퍼셉트론(MLP, Multi-Layer Perceptron)이라고 한다[6]. MLP는 복잡한 문제를 해결하기 위해 깊은 층수를 쌓는데 이 경우, 학습 과정에서 데이터가 사라지는 문제가 발생한다. 또한 ANN은 새로운 사실을 추론하고 새로운 데이터를 처리에도 한계점을 드러냈다. 이런 한계점을 해결한 딥러닝이 2006년에 등장하였다[7].

딥러닝의 알고리즘에는 심층신경망(DNN, Deep Neural Network), 합성곱신경망(CNN, Convolution Neural Network), 순환신경망(RNN, Recurrent Neural Network), 제한된 볼츠만 머신(RBM, Restricted Boltzmann Machine) 등이 있다.

DNN은 복잡한 비선형 관계들을 학습할 수 있어 분류나 회귀분석에 사용된다[8]. CNN은 특징 맵을 생성하는 필터까지도 학습이 가능해 이미지 처리 신호, 컴퓨터 비전 분야에서 응용된다[9]. RNN은 순차적 정보가 담긴 데이터에서 규칙적인 패턴을 인식하고 추상화된 정보를 추출 가능하여 텍스트, 음성, 음악, 영상 적합하다[10]. RBM는 비지도 학습에 활용되는 분류, 선형회귀 분석, 필터링, 차원축소, 필터링 등에서 유용하게 사용되고 있다[11].

## 2. 심층신경망 수행과정

제안하는 연구에서 회귀분석에 적합한 DNN 알고리즘을 사용하여 학습한다. DNN 수행과정은 그림 3과 같이 데이터를 기댓값에 가깝도록 표현 학습하는 방식이다. 연속된 층(Layer)에서 점진적으로 의미 있는 표현을 학습한다[12]. 데이터로부터 학습하는데 얼마나 많은 층을 사용했는지가 모델의 깊이가 된다. 처리되는 상세 내용은 일련의 숫자로 이루어진 층의 가중치에 저장된다[13]. DNN에서는 가설이 얼마나 정확하지를 판단하는 손실함수(Loss Function), 비용을 낮추기 위한 경사하강법(Gradient Descent), 경사하강법을 적용하기 위한 시그모이드(Sigmoid)나 활성화함수(Activation Function), 손실함수가 감소하는 방향으로 가중치 값을 수정해 나가는 역전파법(Back Propagation) 등에 알고리즘을 사용한다[14].

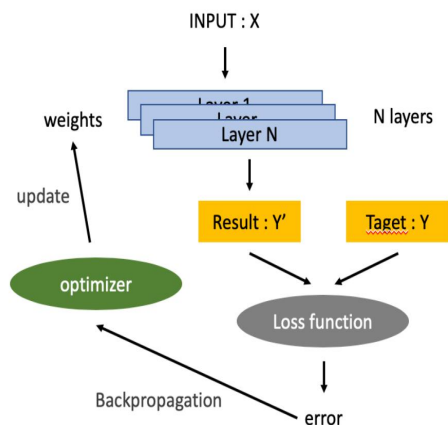


그림 3. DNN 수행과정

## III. 시스템 구성

### 1. 데이터 수집 및 전처리

학습에 필요한 변수가 표 1과 같다. 기상 데이터는 기상청 기상자료개방포털, 농산물 가격 데이터는 aT 농산물유통정보에서 수집한다.

표 1. 가격 예측을 위한 데이터

구분	변수(지역)	기록단위	기간
기상	평균기온(전국)	일간	2010.01~2020.09
	최저기온(전국)	일간	2010.01~2020.09
	최고기온(전국)	일간	2010.01~2020.09
	강수량(전국)	일간	2010.01~2020.09
	평균풍속(전국)	일간	2010.01~2020.09
	최대풍속(전국)	일간	2010.01~2020.09
채소	채소 품목 (전국 소매가격)	일간	2010.01~2020.09

채소 21개 품목을 가격 예측 대상으로 하고, 채소 가격에 영향을 미치는 기상 변인 데이터들을 학습 모델을 만든다. 현재 기준 2010년 1월 1일부터 2020년 9월 30일까지 10년 기간의 데이터들 3,700 건수를 준비한다. 기상 데이터와 다르게 채소 품목가격 데이터들의 경우 공휴일, 주말에는 데이터값이 존재하지 않는다. 기간 중 비어 있는 값들은 전과 후 날짜의 평균값으로 채워 처리한다.

### 2. 시스템 구조

제안하는 시스템은 사용자가 웹 브라우저를 통해 해당 웹 사이트 접속하여 기상 정보를 입력하여 가격 정보를 요청하면 서버는 기상 변인 중심으로 미리 학습된 데이터의 가격 정보를 사용자에게 응답해 주는 방식이다. 시스템 구조는 그림 4와 같다.

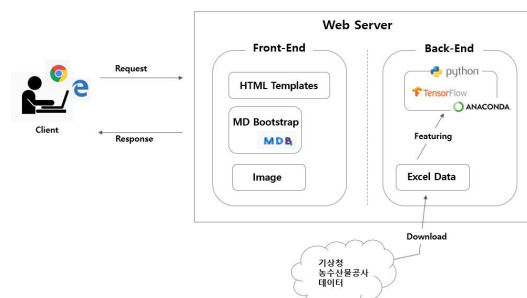


그림 4. 시스템 구조

### 3. 학습 알고리즘

그림 4와 같이 학습 데이터를 다운로드하여 전처리 과정을 거친 후, 학습 모델을 구현한다. 학습 모델 구현을 위해서 가격( $x$ ), 변인의 요소 기온, 강수량, 풍속을 독립 변수( $y$ )로 가설(Hypothesis)을 설정하고 가중치(Weight) 값을 구한다(1). 가설이 얼마나 정확한지 판단하기 위해 예측값과 실제 값의 차이 대한 비용함수(Cost Function)를 구한다(2). 답러닝의 학습은 결국 비용함수를 최소화하여 최적의 해를 구하는 것이다. 이때 최솟값을 구하는 알고리즘 경사하강법 그림 5와 노드에서 다음 노드로 이어지는 가중치를 조절하는 역전파법 그림 6을 구한다.

$$H(x) = Wx + b \quad (1)$$

$$COST(w, b) = \frac{1}{n} \sum_{i=1}^n [y^{(i)} - H(x^{(i)})]^2 \quad (2)$$

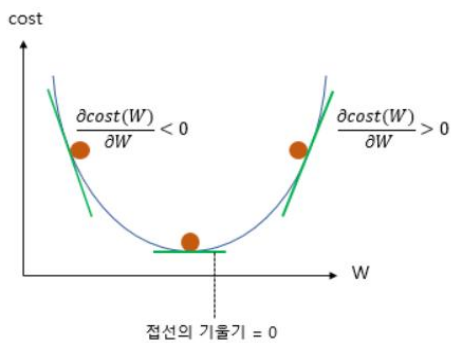


그림 5. 경사하강법

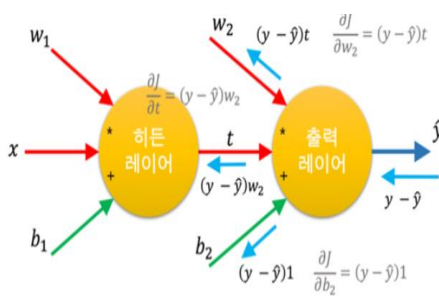


그림 6. 역전파 법

## IV. 실험 및 결과

### 1. 실험 환경

본 논문에서 제안한 가격 예측 시스템의 구성환경은 표 2와 같고, 학습 데이터 셋은 표 3과 같다.

표 2. 가격 예측 시스템 구성환경

구분	설명
Server	Flask Web Server
Back-End	Python, Anaconda, TensorFlow
Front-End	HTML, CSS, Javascript
Framework	MDBootstrap

표 3. 학습 데이터셋

구분	건수 / 총건수
Training	2,590 / 3,700
Testing	1,110 / 3,700

### 2. 실험 결과

본 논문에서 제안한 가격 예측 시스템은 그림 6과 같이 사용자가 웹 사이트에 접속하여 그림 7과 같이 채소 품목을 선택하고 기상 정보를 입력한다. 예를 들어, 채소 품목 배추, 평균 온도 15.5, 최소 온도 3.5, 최대 온도 20.2, 강수량 5, 평균 풍속 2, 최대 풍속 17.5의 정보를 서버에 요청하면 전국 기상과 전국 품목별 채소 소매가격으로 미리 학습된 예측 가격을 사용자에게 그림 8과 같이 응답한다.

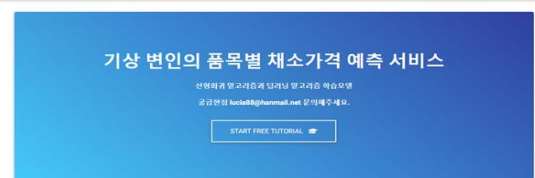


그림 7. 웹 서비스 화면

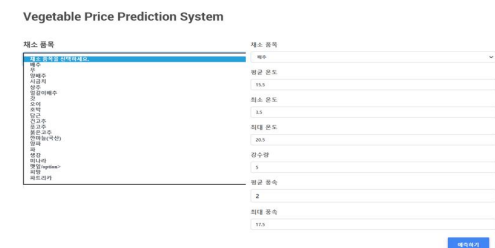


그림 8. 입력 화면

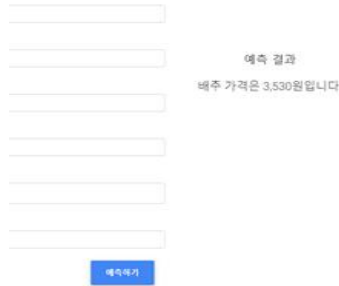


그림 9. 결과 화면

### 3. 검증

본 연구에서는 전체 데이터에서 Train Set 70%, Test Set 30%의 비율로 순차적 분할하였다. 총 학습 횟수 10,000번으로 지정하고 학습률(Learning rate) 0.01로 설정하였다. 비용함수 값의 변화가 거의 없을 때는 Epoch 수가 800~1000회 사이에 가장 좋은 결과를 보였다. 여기서 epoch 수는 학습의 반복을 의미한다. 예측 모델의 정확도 측정 결과, 배추 0.048006, 시금치 0.072450, 오이 0.068229, 갓 0.052364등 예측 정확도가 나타났다. 예측의 정확도는 0과 1사이에 수로 표현된다. 0에 가까우면 예측이 정확도가 높고, 1에 가까우면 정확도가 낮다. 그림 8번은 정확도가 가장 높은 배추 품목의 학습 과정과 결과이다.

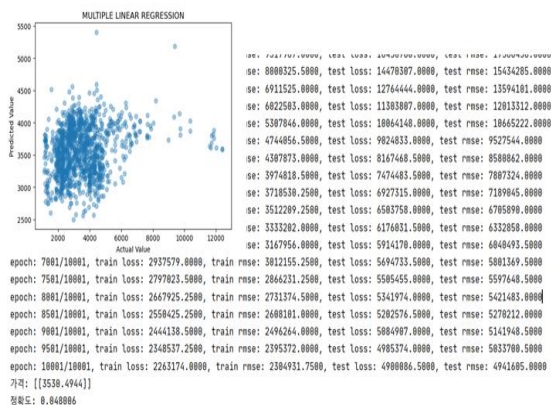


그림 10. 배추 품목 정확도 및 예측 가격

정확도 평가는 실제 값과 예측 값의 차이를 제곱한 평균한 Mean Squared Error 사용하였다(3).

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2 \quad (3)$$

## IV. 결 론

본 연구에 목적은 소비자의 체감 물가에 직접 영향을 주는 채소 가격 변동을 효과적으로 대응할 수 있는 웹 서비스 기반에 가격 예측 시스템을 제안한 것이다. 예측 시스템 구현을 위해 웹 서버를 구축하고 웹 사이트를 구현하였으며, 가격 영향을 미치는 변수들의 10년치 데이터를 수집하고 3,700 건수를 학습에 활용하였다. 학습을 위해서는 다변형 선형회귀와 딥러닝 알고리즘을 사용한 결과, 백분율 기준 92% 예측력을 보였다.

본 연구의 가격 예측 시스템을 통해 생산자는 공급량을 조절하고 소비자는 합리적인 소비에 활용될 수 있을 것으로 기대한다. 향후 물가 상승률, 수확량, 재배면적, 수출입, 정부의 수급관리 변인을 추가하며 좀 더 높은 예측을 도출할 수 있을 것이다.

## 참고문헌

- [1] 연합뉴스 10월 20일, <https://www.yna.co.kr/view/AKR20200929148800530?input=1195m>, 2020
- [2] 최병욱, "농업관측정보활용해 수급안정 도모실행조직 육성필요", 한국농촌연구원 보도자료, 2019
- [3] 고영복, 네이버 지식백과 <https://terms.naver.com/entry.nhn?docId=1521625&cid=42121&categoryId=42121>, 2000.
- [4] 이우리, "시계열 분석과 예측", 탐진, 2013
- [5] 네이버지식백과, 오원석 <https://terms.naver.com/entry.nhn?docId=3578519&cid=59088&categoryId=59096>, 2014.
- [6] T.Mehryar Mohri,1,2 and Afshin Rostamizadeh1, "Perceptron Mistake Bounds", 2013.
- [7] Kunihiko Fukushima NHK Broadcasting Science Research Laboratories, "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position", 1989.
- [8] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," IEEE Trans. PAMI, special issue Learning Deep Architectures, 2013.
- [9] T.Sainath et al., "Convolutional neural networks for LVCSR," ICASSP, 2013.
- [10] Mikolov et al., "Recurrent neural network based language model," Interspeech, 2010.
- [11] 오일석, "기계학습", 한빛아카데미, 2017
- [12] Szegedy, Christian, Alexander Toshev, and Dumitru Erhan. "Deep neural networks for object detection."

- Advances in Neural Information Processing Systems. 2013.
- [13] 안성만 딥러닝 모형과 응용사례, 지능정보연구학회지, 2016
- [14] 민성욱, 딥러닝(Deep Learning)을 이용한 주택가격 예측모형 연구 : 서울시 주택 매매 실거래가를 중심으로, 강남대학교, 2017

판사, 1998), 칼스와 전자상거래(법영사, 1999), 소프트웨어 프로세스(라메르정보기술, 2000), 전자상거래시스템론(법영사, 2001)외 15편의 저서가 있으며, 관심분야는 빅데이터, 클라우드, IO, 전자정부, 핀테크, 비즈니스모델링 등이다.

## 저자소개



좌 민 혜 (Min-Hye Jwa)  
2018년 숭실대학교 정보과학대학원 공학석사를 취득하였다. 기업의 직무교육 및 신입 교육, 과학기술정보통신부 이공계 연수 과정, 고용노동부 NCS 교육, 대학에서 프로그래밍 교육하고 있다. 관심 분야는 인공지능, 빅데이터, 클라우드 등이다.



연 광 제 (Kwang-Je Yeon)  
1999년 2월 광주대학교 전자공학과 학사, 2017년 숭실대학교 정보과학대학원 공학석사를 취득하였다. (주)아이티노매즈 ITO 사업 본부장이며 주요 관심분야는 DBMS, 블록체인, 드론 등이다



이 재 은 (Jae-Eun Lee)  
2001년 고려대학교 경영과학 & MIS 석사를 취득하였다. 현재 넬리지튜브의 대표로 재직하고 있으며, 관심분야로는 인공지능, 클라우드, RPA, 디지털거버넌스, 핀테크 등이다



김 광 용 (Gwang-Yong Kim)  
1984년 고려대학교 공학사 졸업, 1991년 미국 미시시피 주립대 보험수리학 석사, 경영학박사를 취득하였다. 저서로는 Ada언어와 응용(정익사, 1987), CALS/EC (법영사, 1996), 인스턴트코바(영진출