머신러닝을 이용한 양상추 가격 예측 - 변화점 추적 분석을 중심으로 -

김유현*, 박주환*, 윤금비*, 김정우*
*강릉원주대학교 경제학과
e-mail:q03150@gwnu.ac.kr, qkrwnghks821@gwnu.ac.kr,
ykbj7@gwnu.ac.kr, kurtkim@gwnu.ac.kr

Prediction on Lettuce Price Using Machine Learning - Focusing on Detecting Change Point -

Yu-Hyeun Kim*, Ju-Hwan Park*, Geum-Bi Yun*, Jeong-Woo Kim*
*Department of Economics, Gangneung-Wonju National University

요 약

최근 급격한 기후 변화는 각종 농산물의 가격에 영향을 주었다. 앞으로 잦은 기후 변화가 예상됨에 따라 변동하는 농산물의 가격을 예측하기 위해 본 논문은 양상추의 시계열 자료를 주제로 머신러닝을 통해 가격 예측을 진행하였다. 실제 예측값과 근접한 군집을 형성하여 예측을 시도하는 k-평균 클러스터링을 중심으로 예측하였으며, 추가적인 머신러닝 방법을 통해 예측력과 설명력을 높였다. 이러한 방법을 통해 양상추의 가격을 예측한 결과 k-평균 클러스터링이 실제 양상추 가격과 가장 근접한 예측값을 나타냈다.

1. 서론

최근 급격한 기후 변화로 인해 각종 농산물의 가격이 오르거나 수급이 갑자기 중단되는 등의 문제가 빈번하게 발생하였다. 이에 따라, 2020년에는 긴 장마와 태풍의 영향으로 토마토 수급이 어려워지면서 가격이 높이 뛰는 경우가 있었고, 지난해 10월에는 갑작스러운 한파로 채소 가격이 급등하면서 양상추 수급이 어려워지게 되어 양상추 없는 햄버거가 등장했으며 샌드위치 전문점에서는 샐러드 판매를 일시 중단하였다. 이러한 현상들을 보았을 때 앞으로도 기후 변화에 따른 농산물 가격변화가 잦아질 것으로 예상한다.

이에 본 연구는 최근에 타격을 입었던 국내 양상추 가격에 대한 분석과 예측을 통해 소비자 입장 또는 관련 기관들의 농산물 가격 안정화 등에 도움이 되고자 한다. 기존에 이루어진 가격 변동성 측정 연구들에서 사용된 주요 독립변수는 크게 재배면적, 강수량[1], 거리두기 단계[2], 가축의 질병 발생률[3]이 사용되었다. 이를 바탕으로 본 연구에서는 양상추의 가격 변동을 예측하며, 주어진 데이터로부터 새로운 하위 샘플을 추출하여 이용하는 머신러닝 기법인 k-평균 군집화 기법을 활용하여 예측 연구를 수행하였다.

2. 연구 방법

기존의 k-평균 클러스터링 방법의 기본적인 수식은 (1)과 같다.

$$S = \arg_S \min \sum_{i=1}^{K} \sum_{x \in S_i} \| x - \mu_i \|^2 \quad (1)$$

위 식에서 군집들은 각 군집에 속한 관측치가 각 군집의 평균까지의 거리가 최소화되도록 선택된다. 즉, k-평균 클러스터링 방법은 군집 내 분산을 최소화하는 방식을 따르고 있다. k 값은 연구마다 임의로 주어지거나 10개 내외에서 탐색 되는 경우가 많으나 데이터의 크기에 설정에 따라 비효율적이거나 정확한 예측 정확성이 낮아질 수 있다.

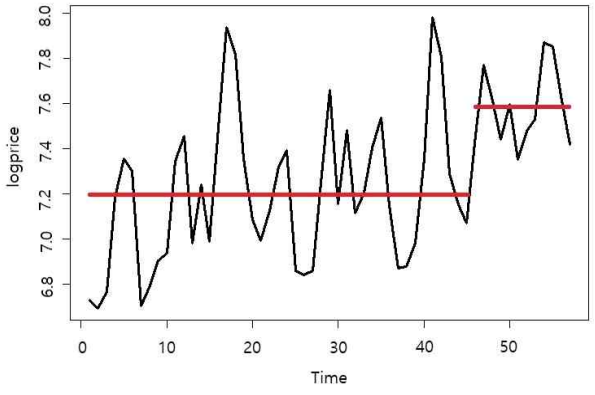
이에 따라, 본 논문에서는 2개에서 8개 사이의 k 값을 설정 군집화하여 정확한 예측을 시도하였다. 추가적으로 k-평균 클러스터링 방법 이외의 머신러닝 기법을 활용하여 예측성능을 비교하고자 하였다. 먼저 선형회귀 분석 기법은 독립변수와 종속변수 간 상관관계를 통해 산출된 예측함수로 예측모형을 만들고 예측을 실행한다. 그러나 이렇게 기본이 되는 예측모형은 과적합[4]이라는 맹점을 지닌다. 따라서 이후 연구에서는 앞선 선형 모델의 예측력과 설명력을 높이기 위해 추

가적인 정규화 방법인 LASSO를 활용해 데이터를 처리했다. LASSO는 영향력이 적은 독립변수에 곱해지는 가중치를 0에 수렴시킴으로써 효율적인 예측을 기대하게 했다. K-NN Regression(K-최근접 이웃 회귀) 기법은 새로운 데이터가 주어졌을 때, 새로운 데이터에 가장 가까운 k개 데이터 평균을 통해 예측한다. 이러한 방식도 과적합 문제가 나타는데 이에 이번 연구는 최적의 k 값을 선택하여 과적합 문제를 해결하였다. 마지막으로 데이터 마이닝에 기반한 모델 의사결정나무 모형 또한 비교에 활용되었다. 이번 연구는 예측함수의 맹점이 되는 과적합 문제를 해결하여 정확성을 높이려는 시도를 지속한 가운데, 가장 정확한 모형은 k-평균 클러스터링 방법일 것으로 예측된다.

3. 분석 결과

양상추의 가격을 예측하기 위하여 독립변수로는 강수량, 양상추 키워드 검색지수[5], 코로나 19로 인한 사회적 거리 두기, 가축 질병 등을 본 연구에서 사용하였다. 데이터 수집기간은 2017년1월부터 2021년 12월까지의 월별데이터이며, 이중에서 2020년 7월부터 2021년 12월까지가 예측 구간이다.

특히, 본 연구에서는 양상추 가격의 시간에 따른 변동을 고려하여, 그림 1과 같이 양상추 가격의 시계열 상 구간평균이 변화하는 지점(Change point)을 추적하여 2개의 구간(A, B)으로 구분한 후 머신러닝 기법을 적용하였다. 이러한 구간 구구분은 시계열 자료의 안정성을 높여, 예측의 정확성을 도모할 수 있다.



[그림 1] 구간분석

여기서 양상추의 가격의 k-평균 군집화 기법의 예측 정확성을 측정하기 위해 절대비오차(MAPE)를 사용하였으며, 비교할 머신러닝 기법으로는 OLS, LASSO, Decision Tree(Tree), KNN을 사용하였다. 데이터의 기술 통계량은 표 1과 같다.

[표 1] 기술 통계량

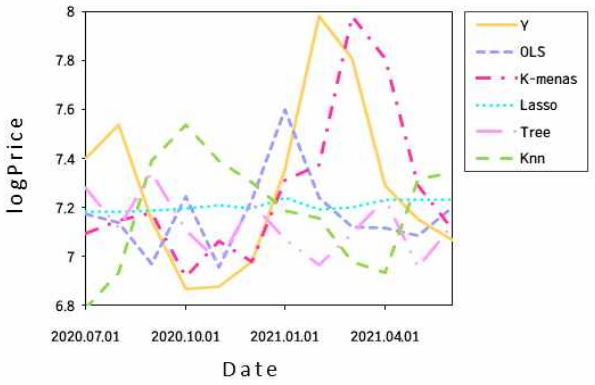
Time	Variable	MEAN	STD	MIN	MEDIAN	MAX
A	Price	1375.6	510.5	728	1270	2927
	pcpt	109.1	103.5	8	65	428
	srch	199.7	64.1	92	179	345
	Sd	0.375	0.84	0	0	3
	Ld	123.8	46.6	50	112	230
B	Price	1994.4	356.9	1557	1859	2624
	pcpt	104.1	86.9	7	76	288
	srch	127.5	54.6	84	113	285
	Sd	3.25	1.2	2	3	5
	Ld	104.75	29.9	69	102	159

*pcpt(precipitation), srch(search), Sd(social distancing), Ld(livestock disease)

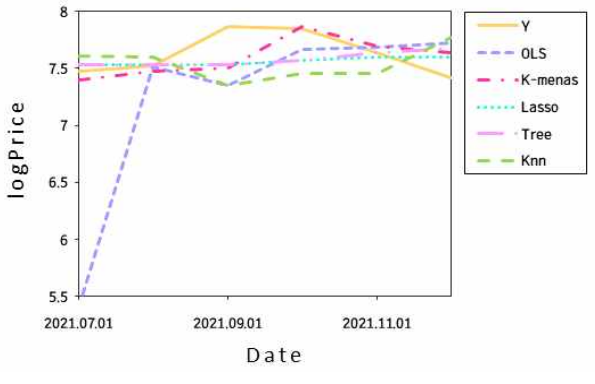
예측 기법들에 따른 예측 성능은 표 2와 같다. k-평균 클러스터링 방법이 가장 낮은 MAPE 값을 보여주며 예측 성능이 가장 높은 것으로 나타났다.

[표 2] MAPE 비교

Tme	OLS	kmeans	LASSO	Tree	KNN
A	3.95	2.77	3.69	4.01	6.3
B	6.91	1.73	1.96	2.05	3.56



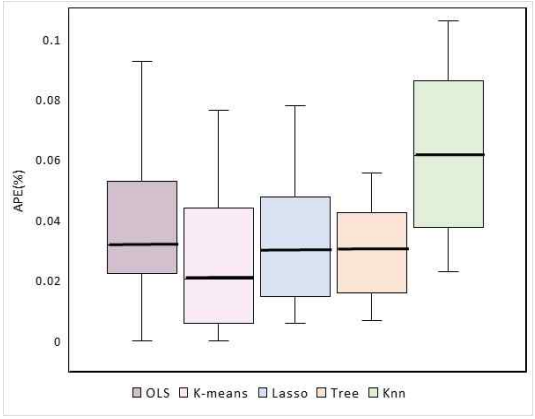
(a) A 구간



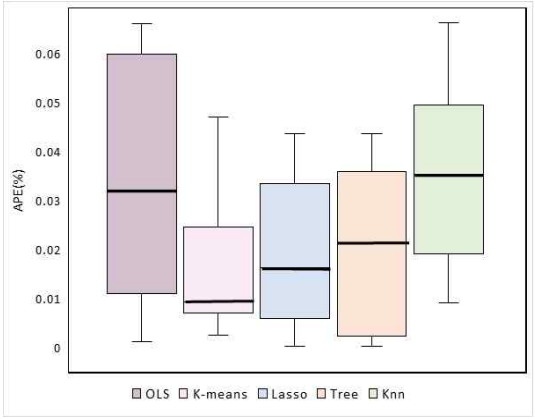
(b) B 구간

[그림 2] 예측 가격의 비교

그림 2는 2개의 구간별로 k-평균 클러스터링 방법과 다른 예측 기법들의 예측 추세를 비교하고 있다. 표 2에서 가장 낮은 MAPE 값을 나타낸 k-평균 클러스터링 방법이 그림 2의 A와 B를 통해서 양상추의 실제 가격(노란색)과 가장 근접한 추세를 보여주는 것이 k-평균 클러스터링(분홍색)임을 더욱 명확하게 보여주고 있다.



(a) A



(b) B

[그림 3] 상자 그림 비교

그림 3은 상자 그림을 통하여 구간별로 예측 기법을 비교한 결과이다. 표 2를 통해 k-평균 클러스터링이 가장 낮은 MAPE를 나타내며 중위값이 가장 낮은 수준으로 나타났으나, B의 상자 그림에서 Lasso나 Tree에 비교해 최솟값과 최댓값이 높게 측정되었음을 알 수 있다. 하지만 극단치들이 존재할 경우 중위값이 대푯값으로 적절하기 때문에 k-평균 클러스터링의 예측 정확성은 유효하다고 볼 수 있다.

4. 결론

본 논문에서는 변화점 추적을 통하여 시계열 자료를 2개의 구간으로 나누어 k-평균 군집화 방법을 통하여 양상추의 가격을 예측하고 k-평균 군집화 방법의 예측 정확성을 입증하였다. k-평균 군집화 방법은 다른 예측 기법보다 낮은 수준의 예측 오차를 보여주며 정확성이 높은 것으로 나타났으며, 실제 양상추 가격과 비교하여 가장 근접한 예측값을 보여주었다. 상당히 안정된 예측 성능을 보이는 k-평균 군집화의 경우 더 높은 단계를 클러스터링을 고려할 수 있는데, 자료의 크기가 상당히 큰 경우 높은 단계의 클러스터링 반복 횟수 간의 분석적 접근이 향후 의미 있는 연구가 될 것으로 보인다.

참고문헌

[1] 장정현, 김지원, 곽다은, “과일 도매가격과 날씨 요인에대한 상관관계 연구”, 한국정보처리학회 학술대회논문집, 제24권 2호, pp. 706-708, 11월, 2017년

[2] 김용준, “식량 공급망의 위기와 농업의 대응”, 정책연구, 13호, pp. 213-236, 8월, 2020년

[3] 국승용, 서홍석, 서동주, 권상욱, 김경진, “최근 농산물 가격 변동 실태와 시사점“, 농정포커스, 201호, pp. 1-25, 10월, 2021년.

[4] 배성완, 유정석, ”머신러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측“, 주택연구, 제26권 1호, pp. 107-133, 2월, 2018년

[5] 노호영, 김성용, 김태영, “인터넷 검색지수는 농식품 구매를 선행하는가?”, 농촌경제, 제42권 제2호, pp. 1-34, 6월, 2019년