

논문 2024-2-4 <http://dx.doi.org/10.29056/jsav.2024.06.04>

기계학습 기반의 감귤 착과량 예측 시스템

이봉규*†

A Prediction System for Amounts of Fruits of a Citrus Tree based on Machine Learning

Lee, Bongkyu*†

요 약

농업기술에 정보통신기술(ICT)을 접목하기 시작하였으며, IoT(Internet of Things) 장치를 이용하여 측정된 온도, 습도, 토양 등에 대한 데이터를 인공지능으로 분석하여 작물 생산을 제어하는 연구가 활발히 진행되고 있다. 감귤은 생육단계에 대한 정보를 디지털 자료화할 수 있어서 정보통신기술을 접목한 생산제어가 가능하다. 본 논문에서는 감귤 착과량과 연관 관계가 있는 데이터를 이용하여 실제 감귤 착과량을 예측하는 새로운 형태의 인공지능 기반 예측 방법을 제안한다. 제안하는 방법은 감귤 착과량을 예측할 수 있는 요인에 해당하는 데이터를 효과적으로 학습할 수 있는 예측 모델을 제시한다. 제시된 모델은 최적의 예측을 위하여 앙상블 형태의 복합 모델로 구성된다. 학습된 예측 모델은 실제 측정된 제주 감귤 데이터를 이용한 예측 성능에 대한 실험을 수행하였다. 평가실험을 통하여 제안시스템이 효과적으로 감귤 생산량 예측을 할 수 있음을 보였다.

Abstract

Information and communication technology (ICT) is starting to be applied to agricultural technology. And Research is being actively conducted to control crop production by analyzing data measured using IoT devices with artificial intelligence technologies. Since information about the growth stages of Citrus can be converted into digital data, production control is possible by information and communication technology. In this paper, we propose an artificial intelligence method that predicts the actual amount of citrus. The proposed method presents a prediction model that can be effectively learn data corresponding to factors that can predict the amount of citrus. The model consists of a complex model in form of an ensemble for optimal prediction. The performance of the proposed model is verified using data obtained from actual citrus trees.

한글키워드 : 착과량, 기계학습, 회귀모델, 앙상블, 예측

keywords : Amount of fruit, Machine learning, Regression, Ensemble, Prediction

* 제주대학교 데이터사이언스학과

† 교신저자: 이봉규(bklee@jejunu.ac.kr)

접수일자: 2024.05.03. 심사완료: 2024.06.10.

게재확정: 2024.06.20.

1. 서 론

제주는 겨울철 대표 과일인 감귤의 99%를 생산하고 있으며 지역경제에서 감귤이 차지하는 비

중이 매우 큰 지역이다. 따라서 제주에서는 감귤이 시장에서 적절한 가격으로 판매되도록 매년 감귤의 생산량을 적정 수준으로 유지하기 위하여 재배 감귤에 대한 생산량 관측 조사를 제주도 조례에 따라서 추진하고 있다. 추진 내용은 매년 5월에 개화량 조사를 시작으로 8월에 착과량 조사 그리고 11월에 수량 및 품질조사를 수행한다. 이중 착과량은 실제 생산량을 추정할 수 있는 가장 핵심적인 자료로 활용되는 중요한 지표로 이용되며, 현재는 조사원들이 일일이 수작업을 통하여 조사하는 방법을 사용하고 있다. 이런 조사 방법은 조사자들의 착과에 대한 기준 불일치로 데이터 신뢰성에 문제가 발생하여 착과량 조사의 정확도를 떨어뜨리고 있으며, 노동력 과부하 및 인력 동원에 따른 고비용 문제도 발생한다, 이러한 기존 조사 방법의 문제점을 해결하기 위해 정보통신기술을 활용한 효율적이고 정확한 착과량 조사 방법이 연구될 필요성이 커지고 있다.

농업 분야에 정보통신기술을 접목하는 연구가 최근 활발히 진행되고 있다. IoT(Internet of Things) 장치를 이용하여 측정된 온도, 습도, 토양 등에 대한 데이터를 인공지능과 같은 최신 ICT 기술로 분석하여 작물 생산을 제어하려는 연구들이 대표적이다[1-3]. 감귤나무는 생물계절코드(BBCH-scale)를 기준으로 감귤의 생육단계에 대한 기준을 디지털 자료화할 수 있어서 정보통신기술을 접목한 생산제어가 가능한 대표적인 작물이다. 감귤 착과량의 경우는 봄철에 관찰되는 새순비, 식재된 위치, 온도, 수령, 나무 크기 및 품종 등의 데이터를 통하여 추정할 수 있다[4]. 즉, 기상환경의 지역적 차이와 나무의 크기에 따른 새순의 생육에 따른 데이터를 분석하면 착과량에 대한 예측이 가능하므로 감귤 착과량의 조사에 정보통신기술을 접목할 수 있다. 더욱이 측정할 수 있는 데이터를 바탕으로 인공지능기술을 활용하여 감귤의 착과량을 예측하는 방법은 기존 조사 방법의

문제점을 근본적으로 해결할 것이다.

본 논문에서는 감귤 착과량과 연관 관계가 있는 데이터를 이용하여 실제 착과량을 예측하는 새로운 형태의 인공지능 기반의 예측 방법을 제안한다. 연구 내용은 감귤 착과량을 예측할 수 있는 요인에 해당하는 데이터를 효과적으로 학습할 수 있는 예측 모델을 구축하고, 구축한 모델을 최적으로 학습시킨다. 학습된 예측 모델은 실제 측정한 제주 감귤 데이터를 이용한 실험을 통하여 성능을 평가하고 분석한다.

2. 기계학습과 생산량 예측

작물의 생산량을 자동으로 예측하는 연구가 현재 활발히 진행되고 있다. 고전적인 통계적인 기법으로는 시계열 분석을 들 수 있다. 생산량에 관련성이 있는 환경에 관련된 다양한 정보를 활용하는 다변량 Time-Series 기법을 통해 생산량을 예측하는 연구가 활발히 진행되었다. 그 예로써 환경 관련 정보를 활용하여 국내에서 생산되는 쌀의 토지 생산성을 예측하는 연구가 진행되었는데, 1946년에서 2017년까지의 데이터를 이용하였고 사용된 데이터는 온도, 강수량, CO₂ 농도와 같은 환경정보를 이용하였다[5]. 또한 Spatial (공간) 정보와 Environment (환경) 정보를 활용하여 국내의 양파 생산량을 예측하는 방법을 연구한 사례도 있다[6].

최근에는 인공지능 모델 중 하나인 인공지능경망에 기반한 방법을 이용하여 재배에 따른 생산량을 추정하는 방법을 연구하였다. 특히 심층신경망 모델인 RNN (Recurrent Neural Network)의 일종인 LSTM(Long Short-Term Memory)이 적극적으로 활용되었다. LSTM의 경우 시간적 데이터에 잘 적용되는 특성이 있고, 전통적인 통계적 기법에 비하여 많은 관련 변수를 사용할

수 있으므로 예측력이 높다는 장점을 이용하여 이 모델을 이용한 미국 주요 곡물인 옥수수의 연간 생산량을 예측한 연구가 있다. 기상정보(강수량, 풍속, 온도 등)와 토양 정보(습도 및 품질) 등을 일 단위의 3차원 벡터로 만들고 이를 이용하여 생산량을 예측하는 방법을 구현하였다[7]. 또한 기존의 생산량 추정 위주의 연구와 달리 시설 농가에서 얻어지는 데이터에 LSTM, SVM (Support Vector Machine), RF (Random Forest) 기법을 적용하여 일별 토마토 생산량에 관한 예측 연구를 진행하였는데, CO₂ 밀도, 시설 내부와 외부의 온도와 일사량을 입력 데이터로 사용한 예측 방법이다[8].

이렇듯 작물의 생산량을 자동으로 추정하는 여러 연구가 진행되었으나 현재의 연구들은 대부분 대규모 지역을 대상으로 한 것이기 때문에 연구 결과를 그대로 개별 농가에 대한 생산량 예측에 바로 적용하기는 어렵다. 또한, 공공기관 등에서 제공하는 대규모 지역에 대한 환경정보를 사용하기 때문에 좁은 특정 지역과의 밀접성이 떨어진다. 이에 본 연구에서는 실제 농가에서 얻어지는 실측 데이터를 활용하는 예측 모델을 이용하여 이러한 문제점을 해결한다.

3. 감귤나무 착과량 예측 시스템

본 논문에서 제안한 예측 모델은 회귀분석 모델을 기반으로 한 앙상블 학습모델이다. 회귀모델에 기반을 둔 앙상블 모델을 선택하는 이유는 감귤의 착과량에 영향을 미치는 많은 요인이 연구되어 있고, 추가적인 연구를 통하여 더 많은 요인이 밝혀지고 있어서 학습모델은 이러한 변화에 적응할 수 있어야 실용성이 있는 것이다. 따라서 데이터의 다양한 변화에 적응이 가능한 회귀 기반의 학습모델은 최적의 선택이다.

최적 예측 시스템 구축을 위해 먼저 감귤 착과량에 영향을 미치는 요인을 선정한다. 본 연구에서는 감귤의 착과량에 가장 큰 영향을 미치는 새순비(전체 잎에 대한 새로운 잎의 비율)를 중심으로 5개의 값(독립변수)을 입력으로 사용한다(표. 1). 일반적으로 감귤 착과량은 봄철에 나타나는 개화비(전체 가지에서 꽃을 개화한 가지의 비율) 및 새순비와 매우 유의미한 상관관계를 보이는 것으로 알려져 있다[2].

표 1. 제안시스템 입출력
Table 1. Inputs/Output of the proposed system

범위		데이터값
입력	범주형	품종, 고도, 온도
	연속형	나무수령, 새순비 (새로운 잎 비), 채고
출력		착과량(man_count)

본 연구에서 성능평가에 사용한 데이터는 인공지능 학습용 데이터 사업을 통하여 획득한 제주도에 있는 감귤나무 5,000그루에서 측정한 새순비를 포함하는 중요 식생 정보로 구성된 데이터베이스에서 얻는다[9]. 이들 중에서 20,000개의 데이터를 추출하여 활용한다(표 2).

표 2. 실험 데이터 사양
Table 2. The specification of experimental data

Data 유형	수량
Train	6,788
Validation	1,698
Test	11,514

실제 모델의 구현은 데이터 전처리, 모델링, 및 하이퍼파라미터 조정 등의 프로세스를 자동화할 수 있는 Python의 AutoML 라이브러리인 Pycaret[10]을 사용한다. 구현 방법은 다음과 같다. 먼저 Train 데이터를 이용하여 Pycaret에서 제공하는 회귀모델 중에서 최적으로 학습이 되는 회귀모델 3가지를 선택하는 과정을 수행한다.

모델의 선택에 사용되는 최적화 평가지표는 평균 절대 백분율 오차 (Mean Absolute Percentage Error) 값을 이용한다. 평균 절대 백분율 오차는 수식(1)에 나타난 것과 같이 오차를 계산한 후, 퍼센트로 바꾸어 준 것으로 $MAPE$ 로 나타낸다. 수식(1)에서 y_i , x_i 는 실제값과 추정값이며 n 은 시행 횟수이다.

$$MAPE = 100 \left(\frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - x_i}{y_i} \right| \right) \quad (1)$$

선정 방법은 학습된 모델에 대해서 $MAPE$ 값이 적은 순서대로 3가지 모델을 선정한다, 선택하는 모델을 3가지로 한 것은 최종 구축되는 예측 시스템의 학습 효율을 고려하면서, Voting의 개념을 도입할 수 있는 최소의 모델 수를 유지하기 위함이다. 선택되는 3개 모델은 Pycaret 라이브러리에 존재하는 모델 중에서 학습데이터를 가장 잘 학습하는 모델로 간주할 수 있다. 최종적으로 $MAPE$ 를 기준으로 선택된 3개의 회귀모델은 앙상블되어 하나의 예측시스템으로 통합된다. 그림 1에서는 최종적으로 구현되는 예측 시스템의 구조를 보여준다. 여기서 h_i 는 회귀모델 출력을 나타내고 α_i 는 회귀모델에 대한 최종 가중치이다,

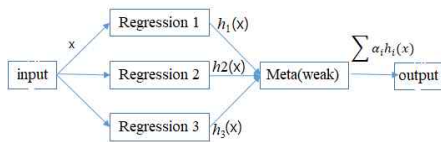


그림 1. 제안된 예측시스템 구조
Fig. 1. The architecture of the proposed system

4. 구현 및 실험 결과

실제 모델에 대한 구현과 실험을 위하여 구성된 실험 환경은 표 3과 같다.

표 3. 실험을 위한 실험 환경
Table 3. The environment of experiments

CPU	Intel(R) CPU 2.60GHz
Memory	204G
GPU	Nvidia Geforce RTX 3090Ti
OS (언어)	Linux (Python 3.8.13)
학습 조건	train_size = 0.7, session_id = 42, normalize = True, silent = True, transformation = True, use_gpu=True, fold_shuffle=True

첫 번째는 평가지표 $MAPE$ 를 최소화로 학습하는 회귀모델 모델 3가지를 선정하는 실험을 표 2에 나타난 train 데이터를 이용하여 실행하였다. 실험 결과 선택되는 3가지 모델은 RF (Random Forest), ET(Extra Trees)와 CatBoost다. 표 4에는 선택된 3개의 모델이 보이는 $MAPE$ 값이 나타나 있다. 표에서 보듯이 3개의 모델이 보이는 값은 모두 $MAPE$ 가 0.1 미만이 되도록 학습할 수 있는 성능을 보였다. 이러한 결과를 바탕으로 볼 때, 이 3가지의 모델이 앙상블 되면 감귤 착과량을 효과적으로 예측할 수 있을 것으로 보인다.

표 4. 모델별 평가지표 값
Table 4. The evaluation indices of models

시스템	MAPE
Random Forest	0.0620
Extra Trees	0.0657
CatBoost	0.0768

선택된 3가지의 모델은 앙상블되어 최종적인 예측 시스템으로 구성된다. 구성된 앙상블 모델은 그림 1에서 보는 것과 같은 형태를 가지며 train 데이터를 통하여 학습한 결과 중요파라미터인 α_i 는 (RF, ET, CatBoost)에 대해서 (0.98, 0.41 0.42)이다. 구현이 완료된 앙상블 모델은 검증데이터 (validation data)를 이용하여 최종적인 조정 및 성능 검증을

수행한다. 이렇게 학습이 완료된 앙상블 모델에 테스트 데이터를 적용한 실제 예측 실험을 통하여 *MAPE* 기준으로 실제 작과량 측정값에 대하여 0.05 이하의 정확도를 내는지를 확인하였다. 실험 결과 제안된 앙상블 모델은 테스트 데이터에 대해서 사용된 성능척도 기준으로 0.0438을 나타내어 0.05 이하로 설정된 기준을 만족하였다. 이 기준에 의하면 제안 앙상블 시스템은 작과량을 실제 값에 5% 이내 기준으로 비교적 정확히 예측할 수 있음을 보였다. 만약 작과량에 연관된 추가적인 데이터를 사용한다면 실제값과 더 유사한 예측 결과를 보였을 것으로 판단된다. 그림 2는 실제 구현된 앙상블 예측 시스템의 실제 내부 실행 과정을 보여준다.

실험을 통하여 제안한 감귤 작과량 예측 시스템은 효과적으로 작과량을 예측할 수 있음을 보였다. 또한 실험 결과를 바탕으로 추가적인 데이터를 활용한 실험을 진행할 경우, 더 우수한 결과를 얻을 수 있음을 확인하였다.

	year	new_leaf_ratio	new_width_avg	chl_avg	area	temp_avg	soil_avg
0	25	1.9	156.5	79.339641	12827	29.899195	21.178738
1	25	2.7	199.0	67.341591	12827	21.845172	24.886982
2	25	2.6	156.5	68.465638	1558	18.322341	64.783898
3	25	2.8	199.0	67.519569	12827	21.897143	25.183193
4	48	3.9	164.5	81.428574	2988	11.292968	64.207611
5	35	3.8	151.5	78.768292	5195	21.988283	27.949818
6	25	1.7	156.5	69.975571	12827	29.815184	17.784481
7	25	1.4	199.0	65.158764	12827	19.717699	19.573837
8	25	2.3	199.0	66.578868	12827	28.494356	21.755323
9	75	5.6	184.6	71.480781	13877	28.861817	73.437044

Check Loss

```
27 # MAPE
28 check_metric(test_data[non_count], pred[label], metric=MAPE)

27 0.0438
```

그림 2. 제안시스템 수행 과정
Fig. 2. The processes of the proposed system

공지능 기반의 생산량 예측 방법을 제안하였다. 제안한 모델은 감귤 작과량과 연관된 5개 범주의 데이터를 이용하는 회귀모델 기반의 앙상블 모델이다. 농가의 재배나무에서 얻은 데이터를 통하여 평가한 결과 감귤의 작과량을 실제값과 비교하여 5% 이내 오차 범위에서 예측할 수 있는 것으로 나타났다. 이러한 결과는 감귤의 생산량을 예측할 수 있는 자동화된 방법을 마련하는 방향을 제시한 것으로 향후 제주의 농업 분야에 대한 인공지능 기술 적용을 위한 중요한 사례가 될 것이다.

인공지능 학습모델을 적용한 시스템을 통해 농작물 생산 예측이 가능함은 여러 가지 연구를 통하여 증명되고 있다. 이런 연구 결과를 바탕으로 전 세계적으로 농업 분야의 인공지능을 활용하기 위한 다양한 생육 관련 데이터를 확보하고 표준화하는 움직임이 활발의 진행 중이다. 이렇게 확보되는 데이터는 인공지능기술에 적용되어 작물의 품질향상과 출하량 조절에 사용됨으로써 농업 경쟁력 확보에 이바지할 것으로 보인다. 특히 개인용 기기인 스마트폰을 이용한 인공지능기술인 온디바이스 인공지능 (On-Device AI) 기술의 발전에 따라서, 스마트폰에 다양한 측정 장치를 연결하여 농업 데이터를 농가에서 직접 수집할 수 있도록 할 경우, 농업 분야에 대한 인공지능 적용은 더욱 확대될 것이다, 이런 추세를 볼 때 본 연구 결과를 바탕으로 더욱 정밀한 농업 분야에 관한 인공지능 연구가 필요한 것으로 보인다.

“이 논문은 2024학년도 제주대학교 교원성과 지원 사업에 의하여 연구되었음”

5. 결론

본 논문에서는 농업 분야의 생산성 향상, 비용·노동력 절감 등을 지원할 스마트농업을 위하여 인

참 고 문 헌

[1] S. Li, S. Peng, W. Chen, X. Lu,

- “INCOME: Practical land monitoring precision agriculture with sensor networks”, Computer Communications, 36 (4), pp. 459-467, 2013, DOI : 10.1016/j.comcom.2012.10.011
- [2] S. Kim, Y. Kim, “A Study on the Application of Machine Learning Algorithm to Predict Crop Production”, Journal of the Korea Academia-Industrial cooperation Society, 22 (7), pp. 403-408, 2021, DOI: <http://dx.doi.org/10.5762/KAIS.2021.22.7.403>
- [3] A. Kaya, A. S. Keceli, C. Catal, H.Y. Yalic, H. Temucin, B. Tekinerdogan, “Analysis of transfer learning for deep neural network based plant classification models”, Computers and Electronics in Agriculture, 147 (1), pp. 70-90, 2018, DOI: 10.1016/j.compag.2019.01.041
- [4] Institute of Agricultural Technology of Jeju, Regulation and Pruning of citrus trees, 2014, <https://agri.jeju.go.kr/>
- [5] D. Chung, D. Han, D., “Evaluation of Forecasting Performance of Rice Yield Models under Climate Change”, Korea Environmental Policy And Administration Society, 26 (4), pp. 197-222, 2018, DOI: <http://dx.doi.org/10.15301/jepa.2018.26.4.197>
- [6] Oh, S., Kim, M, “Predicting Onion Production by Weather and Spatial Time Series Model”, Journal of The Korean Data Analysis Society, 19 (5), pp. 2447-2456, 2017, <https://www.dbpia.co.kr/journal/voisDetail?voisId=VOIS00558530>
- [7] Jiang, Z.,Liu, C., Ganapathysubramanian, B. Hayes, and D., Sarkar, S., “Predicting county-scale maize yields with publicly available data”, Scientific Reports, 10(1), 2020, DOI: 10.1038/s41598-020-71898-8
- [8] Chen, T., Yin, H., Chen, H., Wu., L., Wang, H., Zhou, X., Li, X., “TADA : Trend Alignment with Dual-Attention Multi-Task Recurrent Neural Networks for Sales Prediction”, Proceeding of IEEE International Conference on Data Mining pp. 49-58, 2018, DOI: 10.1109/ICDM.2018.00020
- [9] National Information Society Agency, Citrus Fruit Amount Learning Data, 2022, <https://www.aihub.or.kr>
- [10] M. Ali, “PyCaret: An open source, low-code machine learning library in Python”, 2020, <https://www.pycaret.org>

저 자 소 개



이봉규 (Bong-Kyu Lee)

1995.2 서울대학교 컴퓨터공학과 박사
1996.3-현재 : 제주대학교 교수
<주관심분야> 인공지능, 패턴인식