Journal of KIIT. Vol. 18, No. 1, pp. 1-10, Jan. 31, 2020. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2020.18.1.1

ConvLSTM을 사용한 토마토 생산량 및 성장량 예측 모델에 관한 연구

홍성은*[1], 박태주*[2], 방준일*[3], 김화종**

A Study on the Prediction Model for Tomato Production and Growth Using ConvLSTM

Seongeun Hong*[1], Taeju Park*[2], Junil Bang*[3], and Hwajong Kim**

이 논문은 2019년도 농림축산식품부의 재원으로 농림수산식품 기술기획평가원(첨단생산기술개발사업) 지원사업의 연구결과로 수행되었음(116116-03-1-SB010, 시설원예 생산량 증대 및 경영비 절감을 위한 클라우드 기반 자율제어 시스템 개발) 또한 2019년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019007059)

요 약

스마트 팜의 생육, 생산량의 정확한 예측모델은 스마트 팜의 생산성 향상 및 자동화에 가장 중요한 기술이다. 하지만, 국내의 생육 및 생산량 예측 연구는 연 단위, 월 단위의 생산량 예측연구가 대부분이다. 스마트 팜과 같은 농장 단위 데이터를 사용한 예측 모델 개발연구는 미흡하며, 생산량 예측(추정)을 위한 연구에서는 데이터 기반이 아닌 통계모델을 도출하는 연구가 대부분이다. 그렇기에, 본 연구자는 데이터기반 생육, 생산량 예측에 대한 연구로 스마트 팜 환경에서 발생한 데이터를 사용한 예측모델을 개발하였다. 연구에서는 다중선형회귀, 랜덤 포레스트, 딥러닝(ConvLSTM) 알고리즘을 비교분석하였고, 딥러닝 기법을 적용한 ConvLSTM모델이 개별 농가와 평균 농가의 R^2점수가 가장 높았다. 실험 결과 생산량 예측 모델의 R^2점수는 0.981이고, 생장량 예측 R^2점수는 0.805를 얻었다.

Abstract

The most important technology is the accurate prediction model of the growth and production of smart farms. However, domestic research is largely based on annual and monthly production forecast studies. Predictive model studies using farm unit data are insufficient, and studies for output forecasting (assumption) are being conducted to derive statistical models, not data-based ones. Therefore, the researcher developed a data-based growth and production prediction model using data from smart farm environments. In the study, multi linear regression, random forest and deep learning algorithm (ConvLSTM) were compared, and the ConvLSTM model, which applied deep learning technique, had the highest R^2score for individual and average farmers. The R^2score for the production forecast model was 0.981, and the $R^2$for the growth forecast was 0.805.

Keywords

smart farm, deep-learning, machine learning, sequence data, farm automation

* 강원대학교 컴퓨터정보통신공학과 박사과정
- ORCID[1]: https://orcid.org/0000-0002-7469-2439
- ORCID[2]: https://orcid.org/0000-0003-3199-8881
- ORCID[3]: https://orcid.org/0000-0003-0582-1572
** 강원대학교 컴퓨터정보통신공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-3822-390X

· Received: Nov. 21, 2019, Revised: Jan. 07, 2020, Accepted: Jan. 10, 2020
· Corresponding Author: Hwa-Jong Kim
Dept. of Computer and Communications Engineering, Kangwon National Univ., Gangwondaehak-gil 1, Chuncheon-si, Gangwon-do, Korea
Tel.: +82-33-250-6323, Email: hjkim3@gmail..com

Ⅰ. 서 론

최근 IoT 기술의 발전과 정부의 스마트 팜 확대 정책에 따라 ICT 기술을 접목한 스마트 팜 온실이 국내에 점차 보급되고 늘어나고 있다[1]. 2017년 11월 대한민국의 농림 축산 식품부는 스마트 팜 확산을 혁신성장 핵심 선도사업의 하나로 선정하여, 17년도 기준 온실 4,010ha에서 2022년 7,000ha까지 확장할 계획을 발표하였다. 스마트 팜의 주요 역할은 생육환경 유지관리, 환경정보 모니터링, 자동 또는 원격 환경관리를 가능하게 하는 것이다. 스마트 온실의 경우에는 PC 또는 모바일로 온실의 내부 환경을 모니터링하고 내부 환경을 조절하기 위한 장치의 행동을 원격, 자동으로 제어하여 작물의 최적 생장 환경을 유지 관리하는 것이다.

현재 보급되어 있는 시설원예 스마트 팜은 개별 농가 단위로 온 습도를 중심으로 내부 환경을 제어하고 양액 공급 관리를 하는 수준에 머물러 있다. 최근 빅데이터, 클라우드, 인공지능 등의 4차 산업 혁명 기술이 도입되면서 실시간으로 수집되는 데이터를 분석하여 온실을 관리하는 기술에 대한 연구가 본격적으로 수행되고 있다. IoT에 의해 수집된 정보를 모니터링 할 수 있고, 간단한 제어까지 가능한 1세대 스마트 팜은 현재도 가능하다. 하지만 생산, 정밀, 지능화가 갖춰진 데이터 기반의 2세대 스마트 팜의 경우에는 아직 연구단계에 머물러 있다. 정확한 데이터의 수집이 어렵고, 작물의 생육 조건과 환경 조건의 상관관계를 분석, 예측하여 작물의 재배 기준을 제시하는 알고리즘이나 모델을 개발하는 것이 쉽지 않기 때문이다.

대한민국의 농림수산식품교육문화정보원(농정원)에서는 2014년부터 스마트 팜 정보 공유 시스템을 구축하여 스마트 팜 농가의 데이터를 수집하고 있다. 스마트 온실의 환경 및 제어 데이터는 시스템 연계를 통해 1분 단위로 수집하고, 생육데이터는 1부일 단위로 현장 방문하여 수집되었다. 2018년까지 9개의 품목 240개의 농가 데이터를 수집하는 것을 목표로 하였다[2]. 농정원에서 구축한 스마트 팜 빅데이터의 작물 중 온실 토마토를 선택하였다. 스마트 팜의 기술로 재배되는 토마토 농가는 245개로 집계되었으며 데이터의 정제 및 비어 있는 데이터를 제외하는 작업을 수행하여 학습 데이터를 구축하였다. 데이터에는 생육관련 정보인 측정일, 표본번호, 생장 길이, 화방높이, 줄기 직경, 잎 길이, 잎 폭, 잎 수, 개화 군, 착과 군, 열매 수, 수확 군으로 구성되어 있었다. 환경 정보 같은 경우에는 내 외부 환경 요인을 수집하고 있었으며 그 중 내부 환경 요인을 사용하였다[3].

이와 같이 학습 데이터를 구축한 후 생육량 예측과 생산량 예측을 수행하였다. 예측 모델의 개발을 위해 우리는 다중 회귀 분석, 랜덤 포레스트(Random forest), 딥러닝 ConvLSTM(Convolution 레이어를 추가한 Long term Short Term Memory)기법을 사용하였다. 이 세가지 알고리즘의 성능을 비교하여 가장 좋은 성능을 나타내는 모델을 선정하였다. 성능비교에는 MAE(Mean Absolute Error), RMSE(Root Mean Square Error), R^2(R-squared)을 사용했다. 학습 모델의 성능 평가 기준을 따라 분석한 결과 생산량 예측 모델과 생장량 예측 모델에 딥러닝 기법을 사용한 ConvLSTM 모델이 생산량 예측 모델은 평균 R^2 값이 0.981이었고, 생장량 예측 모델의 경우에도 평균 R^2 값이 0.805로 가장 높았다.

본 논문의 2장에서는 관련 연구와 사용한 알고리즘에 대한 서술한다. 3장의 1절에서는 농정원 스마트 팜 빅데이터를 사용한 학습 모델의 구축에 관하여 서술한다. 2절에서는 학습모델의 구조와 그 방법에 대해서 서술한다. 3절에서는 학습 모델의 성능 평가를 수행하기 위한 방법론을 서술한다. 4장에서는 결론과 활용방안에 대해서 서술한다.

Ⅱ. 관련 연구 및 방법론

2.1 관련 연구

Bashar Alhnaity의 연구에서는 토마토 생산량과 생장량을 모두 예측하는데 서포트 벡터 회귀 SVR (Support Vector Regression), 랜덤 포레스트, LSTM (Long Short Term Memory)모델을 사용하였다. 사용된 데이터는 벨기에 Destelbergen에 위치한 PCS (Ornamental Plante Research Centre) 90m^{2}의 온실에서 측정된 줄기의 직경을 타겟 데이터로 입력 데이터는 온실에 장착된 센서로 Co_2, PAR(Photosynthetic

Active Radiation), 내부 온도를 수집했다. 토마토 생산량 예측을 위해서는 UK Greenhouse farm으로부터 수집된 환경정보 Co_2, 습도, 일사량, 외부온도, 내부온도데이터를 입력 변수로 사용하고, 실제 측정된 생산량을 타겟 변수로 사용하여 토마토 생산량을 예측하는데 학습되었다. 60%를 훈련 데이터로 15%를 검증데이터로 나머지 25%를 테스트 데이터로 사용하였으며, LSTM모델이 MSE=0.002, RMSE=0.047, MAE 0.03으로 모든 경우에 좋은 성능을 보이므로 LSTM을 선택하였다[4].

본 연구와 동일한 방식으로 예측을 수행하는 가장 유사한 논문이라고 볼 수 있다. 이 연구에서 제시하고 있는 성능 비교 그림에 actual 값과 예측 값이 모두 0과 1사이의 값으로 나타나 있을 것을 보았을 때, 오차 측정에 있어 주어진 데이터의 스케일링 변환 없이 그대로 사용한 것이 아니라 학습하기 위해 [0,1]사이의 값으로 스케일링한 데이터를 그대로 사용하여 최종 성능평가 및 그래프를 도출하였기에 지나치게 성능이 좋게 나타난 것으로 보이며 이러한 성능평가로 지나치게 좋은 성능이 도출되었거나 확실하지 않은 성능 평가 결과를 도출했을 것으로 예상된다. 또한 영국과 벨기에의 지역적 차이가 있고 다른 데이터를 분석하여 어떠한 연구 성과가 있는지 알 수 없다. 반면 우리의 연구는 동일한 나라와 지역의 토마토 농장의 생장량 및 생산량을 예측한다는 면에서 가치가 있다[5].

W. C. Lin과 B. D. Hill의 연구에서는 상업용 온실에서 재배되는 파프리카의 주간 생산량 예측을 위한 신경망 모델에 대한 연구를 수행하였다. 캐나다 British Columbia의 2000년부터 2005년까지 총 6년간 파프리카의 주간 생산량을 데이터로 사용하였다. 입력 데이터는 주간 빛 누적 량(KJm^2), 주간 24시간 평균 온도, 이전 주간 생산량 1주 이전~ 3주 이전까지 속성으로 6년간 총 156개의 샘플 데이터를 사용하였다. 학습 모델은 인공 신경망을 사용하였으며, 1주 후, 2주후의 생산량을 예측하였다. 1주 후 예측 성능을 R^2으로 구했을 때, 매년 성능의 차이가 꽤 있지만, 평균 R^2은 0.66, RMSE는 0.25의 성능을 보인다. 또한 2주 후 R^2은 0.59, RMSE는 0.29의 성능을 보인다. 2008년도의 연구로 기존 기계학습 알고리즘과의 비교가 없고, 인공 신경망만을 사용하여 다른 알고리즘과 비교가 없다는 한계가 있다. 또한 생산량은 패턴이 있긴 하지만, 일정하지 않고 값의 변화가 큰 것을 고려하면 꽤 좋은 성능을 얻었다고 말할 수 있다. 인공 신경망 즉, 딥러닝 기술이 더욱 발달한 현재에서는 본 연구보다 더 좋은 성능을 보일 것으로 예상된다[6].

Kim Sung Kyeom의 연구에서는 대한민국에서 농가의 소득원으로 가장 영향이 있는 고추의 생육 예측 및 수량 추정 모델을 개발하였다. 고추의 수량 및 품질은 일조량, 강우량, 기온 등과 같은 기상환경 및 재배지의 토양 조건에 의해 큰 영향을 받기에 이런 조건의 데이터를 수집하여 생육 및 수량을 예측하는 연구를 수행하였다. 국립원예특작과학원의 비 가림 하우스에서 고추를 파종하고 유리 온실의 평균기온인 25℃로 설정하였고, 토양 수분 함량을 30%로 유지하고, 양액은 과채류 전용 양액을 ph6.5 500ml를 매주 주었다. 생육 및 수량 특성에 대한 데이터 수집은 생체 중, 건물 중, 엽 수, 엽 면적을 2주 간격으로 154일 까지 총 12회를 조사하였다. 생산량 조사는 6주부터 2주 간격으로 총8회를 실시하여 데이터를 수집 구축하였다. 생육 모델의 성능 평가로는 MSE(Mean Standard Error)를 사용하였고 R^2 값의 평균 대략 0.98의 모델을 개발하였으며, 생육 모델은 생체 중, 건물 중, 초장 및 엽 면적의 예측 회귀 방정식을 정의하였고, 수량 추정 모델은 해당 재배일의 날씨와 비 가림 시설 내의 평균 온도를 사용하여 추정하였다. 이 연구는 생육 예측 및 수량 추정을 위한 실험 환경을 구축하고 실험 환경에서 발생하는 데이터를 수집하여 모델을 개발한 연구로 실제 농가에서 수집되는 데이터보다 데이터 품질이 높아 좋은 연구결과로 도출 될 수 있었다. 하지만 수량 추정의 연구 부분은 해당 재배 일의 일부 날씨 값을 데이터화 하지 않고 수량을 추정하는 방법을 택하여 연구 데이터의 구축에 미흡함을 드러냈다[7].

Lee Jin Hyoung의 연구에서는 대한민국의 농산물 중 중요한 비중을 차지하는 배추를 연구 대상작물로 삼았다. 봄배추와 가을배추의 정식시기에 따라 실시간으로 측정되는 생육지표 값과 재배기간 중의 기상요소를 기반으로 생육 모델 및 생산량 예측 연구를 수행하였다. 이 연구에서는 배추 작물의 특성

상 생육 및 품질에 영향을 끼치는 기상요소를 활용하였다. 봄배추는 국립원예특작과학원의 비 가림 시설에 2016년 3월 9일부터 4월 7일까지, 가을배추는 8월 9일에서 9월 21일까지 2주 간격으로 파종하였다. 데이터는 기온, 지온, 상대습도, 토양 수분 및 광량을 한 시간 간격으로 측정하였다. 또한 생육 정보는 2주 간격으로 5주씩 생체 중, 건물 중, 엽 장, 엽 폭, 잎 수, 엽 면적, SPAD(엽록소 측정 값) 등의 생육 특성을 조사하였다. 또한 온도 관련 요소인 GDD(Growing Degree Days) 변수를 추가하여 생육 도일의 차이에 대한 회귀분석을 실시하였으며, 시그모이달 회귀 모형을 사용하여 모델을 생성하였다. 모델의 성능은 R^2값으로 평가하였으며, 봄과 가을의 배추 생육 모델의 회귀 성능은 봄배추는 평균, 0.98 가을배추는 0.985의 성능을 보여 좋은 성능을 나타냈다[8].

우리는 총 4개의 관련 연구를 살펴보았고 토마토와 관련된 생장 및 생산량 예측 연구는 거의 찾아볼 수 없었으며, 특히 토마토의 경우에는 국내 연구는 거의 찾아볼 수가 없다. 국외의 연구에는 토마토의 생육 및 생장량 예측 연구는 이미지를 사용한 예측에 대한 연구들이 있었을 뿐이다[9].

이러한 이유는 여러 가지가 있다. 실제 환경에서 데이터를 수집할 때 센서의 오류로 인해 오차가 많이 발생한다. 또한 직접 데이터를 측정해야하지만 해당 농가나 직원이 일정기간 동안 지속적으로 측정하기 어렵다는 것이다. 그렇기에 국외 연구에서는 작물에 카메라를 달아 사람이 직접 측정하는 것이 아닌 기계가 측정한 결과를 사용하는 연구가 진행되는 것으로 고려된다. 앞서 언급한 첫 번째 이유를 해결하기 위해 센서 데이터의 측정을 보정하기 위해 인공 신경망 기법을 사용한 연구도 있었다[10].

관련 연구 조사를 미루어보아 국내 스마트 팜의 보급률 향상을 위한 목표와 달리 토마토 작물에 대한 생장 및 생산량 예측 모델 연구가 미흡하다는 것을 파악하였다. 또한 빅데이터 및 딥러닝 기술의 진보에도 불구하고 다른 연구에서는 통계적 모델을 도출하는 연구에 머물러 있다는 것이 농업 연구 분야의 한계로 고려된다. 본 논문의 연구자는 실제 농장에서 수집되는 데이터(농정원 스마트 팜 빅데이터)에 기반하여 농장 별 데이터 분석에 최신 기술을 적용하여 생장, 생산량 예측 모델을 개발하는 연구를 수행하게 되었다.

2.2 방법론

다중 회귀 분석은 다수의 독립 변수들을 사용하여 종속 변수를 예측할 때 사용되는 알고리즘으로 단순 회귀 분석을 확장한 개념이다. 다중 회귀 분석은 다양한 분야에서 활용되었으며, 종속변수와 독립변수들의 상관관계를 파악하는데 사용되었다. 또한 상관관계를 통해 상관계수를 구할 수 있으며, 이를 통해 통계적 수식을 추정하는데 사용되었다. 다중 회귀 식에 있는 β는 각 독립 변수와의 상관계수를 나타내고, X는 독립변수를 나타낸다. 통계적으로는 독립변수들끼리는 관련성이 없어야 한다는 것을 가정하지만 실제 데이터는 완벽하게 관련성이 없기 힘들어 통계적 가정이 깨지기도 하지만, 실제 현상에 대한 예측 모델 개발 시 우수한 예측 성능을 보이므로 다양한 예측 모델 개발 연구에 활용되고 있다[11].

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k \quad (1)$$

랜덤 포레스트는 다수의 의사결정 나무가 모여 모델을 구성한다는 의미를 담고 있으며, 배깅 기법을 적용하여 모델을 생성하기에 알고리즘의 이름이 랜덤 포레스트가 되었다. 배깅 기법은 N개의 데이터를 N번 복원추출(랜덤추출)하여 추출된 데이터에 대해서 각각의 단일 모델을 형성하고 예측된 결과에 대해서 투표나 평균을 사용하여 최종결과를 도출하는 방식을 말한다. 다수의 의사결정나무의 각각의 예측 결과를 투표나 평균으로 결정하여 일반화를 시켜 의사결정나무의 특성상 나무가 깊게 구성되어 생기는 문제인 과적합을 방지하고 정확도가 높은 모델을 얻을 수 있는 기법이다. 배깅 기법의 효과를 보기 위해서 복원 추출한 데이터 별 독립변수를 랜덤하게 선택하고, 최대로 고려할 트리의 개수를 제한한다[12][13].

RNN은 출력 값의 일부가 입력 값에 다시 포함되는 연속성 개념의 신경 회로망을 의미한다. 이러한 종류의 신경망은 시계열 데이터를 잘 학습하지

만, 장기간에 걸쳐 발생되는 패턴을 인식하지 못하는 문제점이 발견되었다. LSTM과 VGRU(Gated Recurrent Unit) 알고리즘이 연구되어 장기간의 패턴을 인식할 수 있게 되었다[14][15]. LSTM은 가중치뿐 아니라 메모리에 대한 추가 정보를 셀 상태에 저장하고 시계열 패턴의 길이도 조정한다. LSTM에는 keep/forget gate layer가 존재하여 얼마나 많은 양의 이전 메모리를 유지할지 결정한다. 이를 통해 패턴의 기간이 길어 RNN으로 반영되기 힘든 패턴의 정보를 사용자가 선택적으로 기억할 범위를 정해 학습 모델에 정보를 반영할 수 있게 된다. 이러한 이유로 LSTM은 주로 시퀀스나 리스트 구조로 연속되는 데이터를 다루기에 최적화된 구조라고 말할 수 있다. GRU는 LSTM의 입력, 출력, forget 게이트의 총 3개의 입출력 단이 존재하는데 이 중 출력부를 제거하였기 때문에 성능은 비슷하게 유지되지만, LSTM보다 구조가 간단하여 연산량이 줄어드는 장점을 제공한다.

기계학습 기반의 학습 모델과 달리 LSTM을 사용한 딥러닝 모델에 대한 연구는 다방면으로 수행되고 있다. 우리는 time-series 데이터의 forecasting을 위해 LSTM에 컨볼루션(Convolution) 계층을 추가한 ConvLSTM 방식을 취하였다. ConvLSTM은 원래 2차원 데이터에서 특성을 학습하기 위해 제안된 모델이지만, 1차원 시계열 데이터에도 적용하는 연구들이 진행되면서 ConvLSTM 방식을 우리의 데이터에 적용 할 수 있을 것으로 예상하였다. 모델의 초기 학습 시 데이터에서 컨볼루션의 커널의 크기를 시간적인 크기(Window size)로 잘라주면 1일의 패턴, 일주일의 패턴, 한 달의 패턴을 학습하도록 딥러닝 알고리즘에 데이터의 제약 혹은 추가 정보를 줄 수 있다. 우리의 연구에서는 데이터의 발생 주기가 주간으로 이루어져있어 1달(4주)을 1개의 패턴으로 간주하고 그 다음 주의 생장, 생산량을 예측하는 모델을 개발하였다[16][17].

III. 토마토 스마트 팜 데이터 분석 모델

3.1 데이터 및 데이터 정제

우리는 대한민국 농림수산식품교육문화정보원(농정원)의 스마트 팜 데이터를 수집한 스마트 팜 코리아의 빅 데이터를 이용하였다. 스마트 팜 코리아에는 토마토, 파프리카, 오이, 가지, 딸기, 국화, 참외의 7가지 작물에 대한 생육정보, 환경정보, 경영정보가 구축되어 있다. 우리의 연구에서는 토마토 작물에 대한 데이터를 수집하여 분석하였다. 하지만, 실제 데이터를 가져왔을 때 데이터를 그대로 사용하기 어려운 상황들이 있어 다음과 같이 전처리 과정을 거쳤다.

생육정보는 대부분 있었으나 환경정보가 비어있는 경우가 많았다. 또한 생육정보와 환경정보가 동시에 존재해야 하는데 그렇지 않은 경우의 농가 데이터가 많았으므로 전체 245개의 농가 중 15개의 농가의 데이터만을 사용할 수 있었다. 또한 15개의 농가 중에서도 휴작기가 아닌 작기 중에도 비어있는 데이터가 많거나, 빠진 환경정보가 많은 것들을 제외하면 4개의 농가만 남아 4개의 농가에 대한 예측 모델을 개발하게 되었다.

농가의 전체 데이터 중 60%만이 정상적인 값이 있다면 제외하였다. 이 정도로 데이터가 비어있다면 평균값이나 근처의 값으로 빈 데이터를 보충하는 방법도 유효하지 않다고 판단하였다. 또한 작기 별 외부 환경 요인이 매우 다르기 때문에 주어진 데이터를 일관된 조건하에서 분석하기 위해 작기가 최대한 겹치는 기간의 데이터를 선택하였다. 작기 별 설정한 기간은 1작기 2017-03-19~2017-06-11일까지, 2작기 2017-12-31~2018-03-11까지, 4작기 2018-10-14~2018-12-21까지이다. 3작기는 수집된 데이터를 사용할 수 없을 정도로 겹치는 부분이 없었기에 제외하였다.

또한 환경 데이터의 경우에는 생육 데이터와 달리 시간 단위로 발생하였기에 단위 통합을 수행하기 위해 주간의 평균, 최소, 최댓값을 구하여 각 환경 정보에 대하여 생성하였다. 대부분의 환경정보가 일정하게 발생하거나, 오전 오후에 따라 변화를 보이기 때문에 최솟값, 최댓값, 평균값을 사용하였다.

표 1에는 분석하는데 사용할 데이터의 속성을 나열하고 있으며, Open-API에서는 더 많은 정보를 제공하고 있으나, 데이터가 비어있는 경우가 많아 사용하지 못한 속성을 제외하였고, 각각의 환경정보는 최솟값, 최댓값, 평균값을 가지고 있다. 선정된 5개

는 대한민국 충청남도 부여군에 위치한 농가들로 유사한 지역 및 환경 조건을 가지고 있다.

표 1. 토마토 농장 학습 데이터의 속성
Table 1. Properties of tomato farm data

Data category	Code name	Description
Growth information	measDate	Date of Measurement (Index)
	growLength	Length of crop growth
	flowerTop	The height at which crops bloom
	StemDiameter	Diameter of crop stem
	LeavesLength	Crop's leaf length, from start to end of long leaves
	LeavesWidth	The leaf width of the crop is opposite the leaf length
	LeavesNum	Leaf number of crops
	flowerPosition	Blooming group conversion value of crop
	fruitsPosition	Fruitimg group conversion value of crop
	fruitsNum	The number of fruit in a crop
Environmental information	EO/FG/TE	The outside temperature (℃)
	EI/FG/CI	The inside carbon dioxide (CO_2)
	EI/FG/HI	The indide humidity (%)
	EI/FG/TI	The inside temperature (℃)
	EL/FG/EL	Soil salinity (dS/m)
	EL/FG/HL	Soil humidity (%)

학습 데이터는 주간 생장량 및 생산량, 환경정보가 포함되어 있는 데이터로 구성하였으며, 전체 데이터의 80%로 테스트 데이터는 전체 데이터의 20%로 정하고 학습을 진행하였다. 입력 데이터는 A 농가 140개 B농가 164개, C농가 188개, D농가 164개, 전체 농가 656개이다. 학습할 때 사용한 특성은 27개로 위의 표 1에 제시된 학습 데이터 속성에서 measDate는 인덱스로 사용했기에 제외했으며, 각각의 환경정보는 3개의 속성을 학습하는데 사용하였다. 학습 데이터는 생육 정보 9개, 환경정보(6 X 3=18개)로 구성되었다.

3.2 작물 생장 및 생산량 모델

우리는 토마토 데이터를 사용하여 생산량(Fruits number)과 생장량(Leaves length) 예측 모델을 학습하고 성능 평가를 수행함에 있어 더 좋은 모델을 판단하기 위해 총 3가지의 알고리즘을 사용하였다. 첫 번째 모델로 통계기반의 학습 모델이며, 다변수의 가중치를 반영하는 다중 회귀 알고리즘을 사용하였다. 기존의 적은 데이터 셋의 예측 모델에 주로 활용되는 통계적 기법이다. 두 번째 모델로 의사결정 나무를 기반으로 weak한 여러 개의 모델을 생성하고 생성된 모델의 평균을 사용하여 일반화가 잘 되어 높은 정확도를 얻을 수 있는 학습 모델인 랜덤 포레스트 알고리즘을 사용하였다. 랜덤 포레스트 알고리즘은 성능 평가로 순위를 매기는 Kaggle 데이터 분석 대회에서 고성능 모델을 개발할 때 주로 사용된다. 세 번째 모델로 딥러닝 기법의 LSTM 에 CNN 계층을 추가하여 과거 데이터를 기억하되 원하는 주기의 패턴을 인식하는 ConvLSTM 모델을 사용하였다. 시퀀스 데이터의 학습을 위해 LSTM 또는 GRU만을 사용했을 때 연구자의 생각보다 성능이 좋지 않았고, 복합적인 패턴이 발생하거나 데이터의 양이 많은 경우에 데이터의 일부 특징을 추출하기 위해 활용되는 ConvLSTM기법을 적은 수의 학습데이터이며, 1차원 데이터(시퀀스 데이터)를 입력으로 주었을 때 어느 수준의 성능을 얻을 수 있는지 확인하기 위해 사용하였다.

연구에 사용된 모델 중 MLR과 랜덤 포레스트는 입력 데이터(1주)를 가지고 다음 주의 생장 및 생산량을 예측하도록 LSTM은 기간을 기억할 수 있으므로 4주의 데이터를 사용하여 다음 주의 생장, 생산량을 맞출 수 있도록 예측하는 모델을 만들었다.

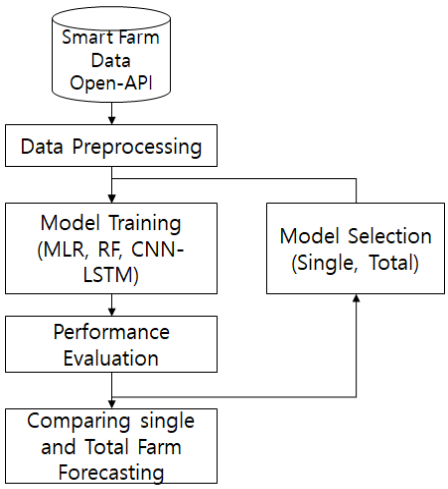


그림 1. 모델 학습 및 비교 분석 흐름도
Fig. 1. Model training and comparative analysis flowchart

그림 1에는 Smart Farm Data Open-API에서 데이터를 가져 온 후 3.1절의 데이터 정제 과정을 거쳤고, 데이터를 0~1사이의 값으로 스케일링 하였다. 학습 및 테스트 데이터를 생성한 후 데이터를 다중 회귀분석, 랜덤 포레스트, ConvLSTM모델로 학습하고 성능 평가를 수행하였다.

그림 2와 같이 ConvLSTM의 모델의 경우 3개월 단위의 재배 기간을 1작기로 고려하기 위해 전체 크기를 3개월(12개의 데이터)을 하나의 입력 데이터로 고려하였다. ConvLayer에서는 필터크기와 커널(Kernel) 크기를 입력으로 주어야하는데 우리는 필터크기를 1로 커널 크기를 4로 설정하였다. 또한 시퀀스 데이터 학습을 위해 그림에 표현되어 첫 번째 사용된 4개의 데이터를 제외한 후 다음을 학습하는 것이 아니라 첫 번째 데이터만을 제외하고 그다음 4주, 그다음 4주 이런 식으로 학습되게 되어있다. 4주로 기간을 잡은 이유는 데이터가 주단위로 수집되었기에 학습할 데이터의 개수가 적은 문제와 동반되어 8주~12주의 재배 기간으로 학습하였을 경우, ConvLSTM 모델이 제대로 학습이 되지 않았고, 4주로 하였을 경우 최적의 성능과 모델의 학습이 이루어졌기 때문이다.

3.3 성능 평가

MAE, RMSE 및 R^2를 사용하여 이러한 예측 모델의 성능을 평가했다. 이러한 평가 측정의 공식은 식 (2)~(4)에 나타내었다. A_t는 실제 값이고 $\hat{A}_t$는 예측 된 값이며 $\overline{A_t}$는 평균값이다.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - \hat{A}_t}{A_t}\right| \tag{2}$$

$$SE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{A_t - \hat{A}_t}{A_t}\right)^2} \tag{3}$$

$$R^2 = \frac{\sum_{t=1}^{n}\left(A_t - \hat{A}_t\right)^2}{\sum_{t=1}^{n}\left(A_t - \hat{A}_t + \hat{A}_t - \overline{A_t}\right)^2} \tag{4}$$

R^2이외에도 Adjusted R^2이라는 평가 지표가 존재하지만 독립변수의 수가 커지면 독립변수가 유의하든 유의하지 않든 결정계수(성능)에 영향을 주는 것을 개선하기 위해 제안된 방법이다. 주로 변수 선택법으로 분석되어 다른 독립변수의 수를 가지는 회귀 모델들의 성능 평가에 사용된다. 본 연구에서 비교 분석하는 모델들이 독립 변수의 수를 변화시키지 않고, 통계적 모델끼리 비교하는 것이 아니라 딥러닝, 의사결정나무와 비교하는 것이므로 Adjusted R^2을 사용하지 않았다.

또한 랜덤포레스트의 하이퍼 파라미터는 test_size =20%, random_state=43, n_estimators=100, min_samples_split=2, min_samples_leaf=1, criterion='mse', max_features ='auto'로 설정하였다.

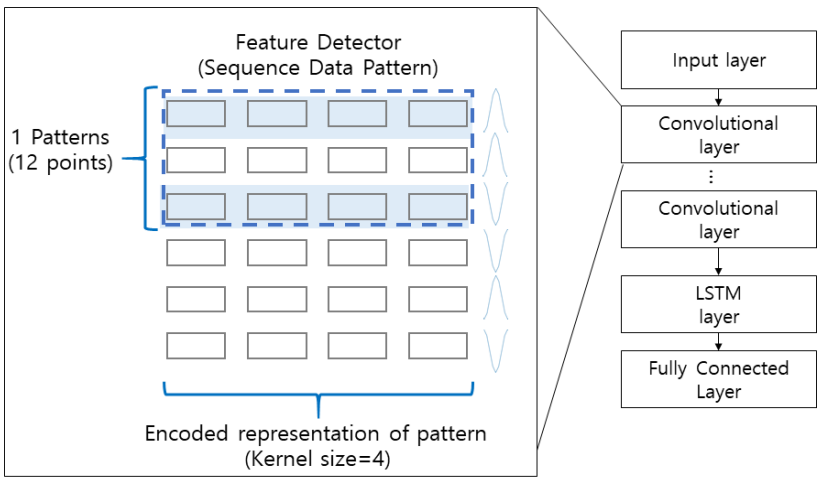


그림 2. ConvLSTM 모델 학습 흐름도
Fig. 2. ConvLSTM model training flowchart

ConvLSTM모델들의 하이퍼 파라미터는 epochs=500, batch_size=64, lr=0.001, lossfunction ='mse'로 설정하였다. 하이퍼 파라미터는 Grid Search 과정을 통해 여러 번의 실험을 통해 설정한 값이다.

표 2에는 생산량 예측 모델의 성능 비교결과를 나타내었다. 표 2의 생산량 예측 모델의 경우 대체로 RF와 ConvLSTM 모델의 R^2점수가 높았다. 가장 높은 R^2 점수를 비교했을 때 RF가 5번 중 3번 성능이 좋았다.

표 2. 생산량 예측 모델 성능 비교표(FruitsNum)
Table 2. Product forecast model performance comparison

Farm	Model	MAE	RMSE	R^2
Total Farm	MLR	11.144	15.220	0.691
	RF	4.161	7.655	0.932
	ConvLSTM	3.588	4.559	0.978
Farm A	MLR	1.162	1.807	0.934
	RF	0.268	0.488	0.995
	ConvLSTM	2.178	2.760	0.987
Farm B	MLR	1.005	1.261	0.946
	RF	0.218	0.400	0.996
	ConvLSTM	0.464	0.542	0.992
Farm C	MLR	3.638	4.445	0.726
	RF	1.881	2.803	0.865
	ConvLSTM	0.742	0.919	0.978
Farm D	MLR	1.162	1.807	0.934
	RF	0.268	0.488	0.995
	ConvLSTM	0.750	0.990	0.972

표 3. 생장량 예측 모델 성능 비교표(leavesLength)
Table 3. Growth forecast models performance comparison

Farm	Model	MAE	RMSE	R^2
Total Farm	MLR	2.871	3.491	0.802
	RF	2.390	3.902	0.741
	ConvLSTM	2.780	3.659	0.792
Farm A	MLR	2.252	3.284	0.852
	RF	2.913	3.715	0.822
	ConvLSTM	2.003	2.697	0.881
Farm B	MLR	1.627	2.052	0.883
	RF	2.523	3.398	0.748
	ConvLSTM	1.301	1.943	0.884
Farm C	MLR	2.261	2.946	0.871
	RF	2.239	2.881	0.781
	ConvLSTM	3.154	4.073	0.771
Farm D	MLR	2.331	3.196	0.576
	RF	2.179	2.766	0.631
	ConvLSTM	1.796	3.019	0.698

하지만 5번의 평균 성능을 비교했을 때, RF는 0.957, ConvLSTM은 0.981로 평균 정확도에서는 ConvLSTM이 약 2.4정도의 더 높은 성능을 보인다고 판단하여 해당모델을 선택하였다.

표 2와 3에서의 성능 비교에서는 학습하기 전 데이터 전처리 과정에서 입력 데이터의 스케일링을 재 스케일링하여 원본 데이터로 환원하여 구한 오차 값이다. 연구자가 지적한 Bashar Alhnaity의 스케일링 문제를 본 연구자는 재 스케일링(환원)과정을 거쳐 오차를 측정하였기에 관련 연구보다 더 정확한 오차 및 성능 평가를 할 수 있도록 하였다.

표 3에는 생장량 예측 모델의 성능 비교 결과를 나타낸다. 생장량 데이터는 생산량 데이터에 비해 일련의 패턴을 찾기 힘든 데이터였기 때문에 비교하는 모델들의 성능이 전반적으로 하락한 것을 확인할 수 있었다.

ConvLSTM 모델의 성능이 총 5번중 3번이나 성능이 높았고, 나머지 2번은 MLR모델의 성능이 좋았다. 5번의 평균 성능을 비교 했을 때, MLR은 0.798의 평균 R^2점수를 ConvLSTM은 0.805의 평균 R^2점수를 가졌다. 미세하지만 전반적으로 MAE와 RMSE도 같이 고려했을 경우에 ConvLSTM의 모델의 성능이 더 높았다.

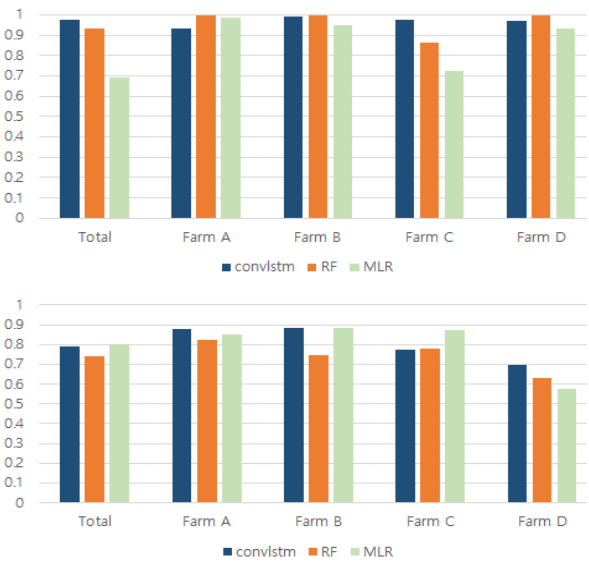


그림 3. 농장 분류와 모델에 따른 R^2 점수 그래프 (상단그래프는 FruitsNum, 하단그래프는 LeavesLength)
Fig. 3. R^2 scores by a farm type & model (Fruits Num, LeavesLength)

그림 3의 상단에는 FruitsNum예측 성능, 하단에는 LeavesLength예측 성능을 나타낸다. 두 예측 성능 모두 ConvLSTM이 가장 좋은 것을 알 수 있다.

Ⅳ. 결론 및 향후 과제

우리는 공개된 스마트 팜 농장 재배와 관련된 데이터인 생육 및 환경 데이터를 사용하여 생산량 및 생장량을 예측하는 모델을 개발하였다. 모델 개발에 있어 총 3개의 학습 모델의 성능을 비교 측정하였다. 생산량 예측에 평균 R^2점수 0.981, 생장량 예측에 평균 R^2점수 0.805의 성능을 보이는 ConvLSTM 모델을 사용하기로 결정하였다.

우리의 연구에서 개발한 모델을 사용한다면, 개별 농가의 생산량 및 생장량을 예측할 수 있고 현재 생육 및 환경 상태를 전체 농가와 비교할 수 있을 것이다. 본 연구의 한계로는 일반적인 시퀀스 데이터의 특성인 짧은 시간동안 발생하는 많은 데이터를 샘플링하여 학습하는 방식인 ConvLSTM의 1c 차원 모델을 사용하여 학습 모델을 구성하였지만, 본 연구에서는 학습에 사용할 데이터가 적었다는 것이다. 학습데이터 구축 시 일 단위로 데이터가 측정되었다면 본 연구자가 사용한 학습 모델을 사용했을 때 더 많은 효과가 있었을 것으로 기대된다. 특히, 농업 데이터는 수집에서 많은 문제가 발생하는데, 데이터 수집 과정이 복잡하고 귀찮다고 생각하기 때문에 실험적으로 구성한 농장이 아닌 일반 농가에서 수집된 데이터의 일관성이 부족하다. 농업 빅 데이터 구축 시 이러한 문제가 항상 발생하고 분석에 사용한 스마트 팜 빅데이터 또한 발생했다고 말할 수 있다. 분석에 사용할 수 있는 데이터가 더욱 많았다면, 스마트 팜의 목적인 생산성 향상에 본 연구의 모델이 영향을 주었을 것으로 예상된다.

추후 연구로는 개발된 모델을 적용하여 개별 농가와 전체 농가의 비교하는 알고리즘을 개발하여 스마트 팜 사용자가 주변 농가 대비 얼마나 잘 재배하고 있는지를 판단할 수 있는 기능을 구현하는 것이며, 또한 수집된 환경정보를 사용하여 최적의 환경과 현재 환경의 차이가 어떠한지를 비교할 수 있는 추가 기능을 구현하여 스마트 팜 농장주가 다른 농가와 비교를 통해 의사 결정 하는데 도움을 주고, 이를 통해 생산성을 향상시키는데 기여하는 시스템을 개발할 수 있을 것이다.

References

[1] T. W. Kim, "ICT-based Smart Farm Greenhouse Status and Outlook", Proceedings of Symposium of the KICS, Vol. 36, No. 3, pp. 256-257. Feb. 2019.

[2] S. H. Shin, M. K. Lee, and S. K. Song, "A Prediction Model for Agricultural Products Price with LSTM Network", Journal of the Korea Contents Association, Vol. 18, No. 11, pp. 416-429. Nov. 2018.

[3] Smart Farm Korea net, https://www.smartfarmkorea.net/front/open/list.do?menuId=M01060201 [accessed: Dec. 31, 2019]

[4] Alhnaity, Bashar, et al, "Using Deep Learning to Predict Plant Growth and Yield in Greenhouse Environments", arXiv preprint arXiv:1907.00624, Jan. 2019.

[5] Salazar, R., López, I., et al, "Tomato yield prediction in a semi-closed greenhouse", XXIX International Horticultural Congress on Horticulture: Sustaining Lives, Livelihoods and Landscapes (IHC2014): 1107. pp. 263-270, Dec. 2015.

[6] Lin, W. C. and B. D. Hill, "Neural network modelling to predict weekly yields of sweet peppers in a commercial greenhouse", Canadian Journal of Plant Science, Vol. 88, No. 3, pp. 531-536, Jan. 2008.

[7] S. K. Kim, J. H. Lee, and H. J. Lee, et al, "Development of Prediction Growth and Yield Models by Growing Degree Days in Hot Pepper", Protected Horticulure and Plant Factory, Vol. 26, No. 4, pp. 424-430, Oct. 2018.

[8] J. H. Lee, H. J. Lee, and S. K. Kim, et al, "Development of Growth Models as Affected by Cultivation Season and Transplanting Date and

Estimation of Prediction Yield in Kimchi Cabbage", Protected Horticulture and Plant Factory, Vol. 26, No. 4, pp. 235-241, Oct. 2017.

[9] Qaddoum, Kefaya, E. L. Hines, and D. D. Iliescu, "Yield prediction for tomato greenhouse using EFuNN", ISRN Artificial Intelligence 2013, Jan. 2013.

[10] Dariouchy, A., et al., "Prediction of the intern parameters tomato greenhouse in a semi-arid area using a time-series model of artificial neural networks", Measurement, Vol. 42, No. 3, pp. 456-463, Apr. 2009.

[11] Amral, N., Ozveren, C. S., and King, D, "Short term load forecasting using multiple linear regression", 2007 42nd International universities power engineering conference, IEEE, pp. 192-1198, Sep. 2007.

[12] Y. Oh, H. Kim, J. Yun, and J. S. Lee, "Using Data Mining Techniques to Predict Win-Loss in Korean Professional Baseball Games", Journal of Korean Institute of Industrial Engineers, Vol. 40, No. 1, pp. 8-17, Feb. 2014.

[13] Y. G. Kim and S. Jong, "Comparing Techniques for Rare Class Classification Problems Using the Overlapped Class Intervals of Data", Journal of Korean Institute of Information Technology, Vol. 8, No. 2, pp. 137-144, Feb. 2010.

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

[15] J. Chung, C. Gulcehre, and K. Cho, et al, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", arXiv: 1412.3555, Dec. 2014.

[16] Xingjian, S. H. I. and Chen, Z., et al, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting", Advances in Neural Information Processing Systems, pp. 802-810, Dec. 2015.

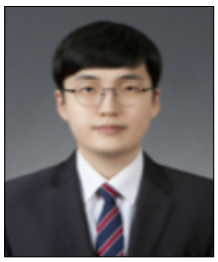
[17] J. Brownlee, "Deep Learning for Time Series Forecasting: Predict the Future with MLPs CNNs and LSTMs in Python", Machine Learning Mastery, Aug. 2018.

저자소개

홍 성 은 (Seongeun Hong)

2015년 : 강원대학교 컴퓨터정보통신공학부(공학석사)
2015년 ~ 2019년 : 강원대학교 컴퓨터정보통신공학과(공학박사) 수료
관심분야 : 빅데이터, 데이터 마이닝, 기계학습, 딥러닝

박 태 주 (Taeju Park)



2015년 2월 : 전남대학교 사회학과(문학사)
2018년 9월 ~ 현재 : 강원대학교 컴퓨터정보통신공학(석사과정)
관심분야 : 빅데이터, 머신러닝 알고리즘, 자연어처리, 감성 분석

방 준 일 (Junil Bang)

2018년 8월 : 강원대학교 컴퓨터정보통신공학과(공학사)
2018년 9월 ~ 현재 : 강원대학교 컴퓨터정보통신공학과(석사과정)
관심분야 : 데이터분석, 머신러닝

김 화 종 (Hwajong Kim)

1984년 3월 : KAIST 전기및전자공학과(공학석사)
1988년 3월 : KAIST 전기및전자공학과(공학박사)
1988년 3월 ~ 현재 : 강원대학교 컴퓨터정보통신공학과(정교수)
관심분야 : 데이터 통신, 컴퓨터네트워크, 네트워크 프로그래밍, 빅데이터