MACHINE LEARNING APPROACHES FOR CROP YIELD PREDICTION WITH MODIS AND WEATHER DATA

Sungha Ju (1), Hyoungjoon Lim (1), Joon Heo *(1)

[1] Yonsei University, 50 Yonsei-ro, Seodamun-gu, Seoul, 03722, Korea
Email: jsh4907@yonsei.ac.kr; joony729@yonsei.ac.kr, jheo@yonsei.ac.kr

KEY WORDS: Crop yield prediction, Machine learning, Neural network, MODIS

ABSTRACT: For accurate prediction, many studies have been actively conducted to estimate grain crops using machine learning techniques. However, there are only few studies which compare the accuracies using many kinds of machine learning techniques for different types of crop. This study was conducted to estimate corn and soybean yields in Illinois and Iowa in the U.S. through four kinds of machine learning techniques, including deep learning algorithms. ANN (Artificial Neural Network), CNN (Convolutional Neural Network), SSAE (Stacked-Sparse AutoEncoder), and LSTM (Long-Short Term Memory) were used as prediction models, and total 14 years of MODIS (MODerate resolution Imaging Spectroradiometer) data, climatic data and crop yield statistics were used as input variables with the six different periodic scenarios. The accuracies were compared in terms of %RMSE (percentage Root Mean Square Error) and compared to the baseline prediction model, which is DT (Decision tree). As the results, CNN model was most accurate over all with the lowest average %RMSE errors in corn (9.36%) and SSAE model in soybean (10.05%) respectively. The best periodic scenario for corn yield prediction was May to September, and for soybean was June to August. This study identified suitable scenario and prediction technique for corn and soybean yield, which will be useful for farming activities and agricultural planning.

1. INTRODUCTION

Accurate yield predictions on grain crops are an important issue in the agriculture planning field and national food security. However, crop yield depends on various factors, such as soil, topography, and management (Chlingaryan et al., 2018), so it is difficult to predict accurately. Since the crop yield has spatially and temporally nonlinear characteristics (Liu et al., 2001), simple statistical analysis methods contain large prediction errors. In order to consider these characteristics, various studies have been conducted to estimate crop yield using meteorological data and satellite images with various vegetation indexes (Gandhi et al., 2016, Chen & Jing., 2017, Ma et al., 2018). Also, many studies have been carried out to estimate crop yield by combining vegetation indexes and meteorological data (Kuwata & Shibasaki., 2016, Ma et al., 2018).

On the other hand, since crop yield prediction have spatio-temporal characteristics as mentioned above, machine learning techniques have been proposed. Decision tree based models (Johnson., 2014, Everingham et al., 2016), artificial neural networks (Chen & Jing., 2017, Fernandes et al., 2017), and deep learning models (You et al., 2017) were used as machine learning based crop yield prediction models. However, there are only few studies to compare the accuracies using many kinds of machine learning techniques for different types of crop yield prediction (Kaul et al., 2005, González Sánchez et al., 2014). In this study, four kinds of machine learning techniques were applied to find the best prediction models for corn and soybean in Illinois and Iowa.

2. METHODOLOGY

2.1 Data

MODIS data, climatic data, and yield statistics were used for yield prediction of corn and soybean in Illinois and Iowa. Four vegetation index images of Normalized difference vegetation index (NDVI), Leaf Area Index (LAI), Enhanced Vegetation Index (EVI), and Fraction of Absorbed Photosynthetically Active Radiation (FPAR) from MODIS satellites from 2003 to 2016 were collected (USGS MODIS., 2019). NDVI and EVI were from MYD13, and LAI and FPAR were collected from MYD15. The weather data was collected from National Climatic Data Center (NCDC., 2019). It contains day-to-day climatic information, which are maximum, minimum, average temperature, precipitation and solar radiation. Yield statistics with county level were collected from National Agricultural Statistics Service of the United States Department of Agriculture (USDA NASS., 2019), and the unit was bushel per acre.

In order to fit all input data to the county level, preprocessing was performed. MODIS data and weather data were stacked into points which comes from land cover maps to extract five weather variables and four vegetation index variables. Then, all the variables were extracted onto county level polygon, and yield statistics were combined. Also, data normalization process was performed to make all the variables have a value range between -1 and 1. Lastly, these data were divided into six periods to compare which period yielded the best prediction results.

2.2 Machine learning techniques

Artificial Neural Networks (ANN): ANN is computational model based on the brain structure and is widely used for machine learning and pattern recognition. ANN was also widely used in crop yield prediction studies (Chen & Jing., 2017, Fernandes et al., 2017). Since the parameters of neural networks including ANN affect the model performance, this study conducted an experiment for parameter optimization. The back-propagation method was scale conjugate gradient, the cost function was mean squared error and the learning rate was fixed to 0.01. The optimal prediction is derived by adjusting the number of hidden neurons in the two hidden layers and epoch size.

Convolutional Neural Networks (CNN): CNN is an improved model of ANN and is widely used as a classification and prediction technique in image segmentation, classification, and recognition. This study used a 2D image map generation method (Ma et al., 2016), using time series of nine variables to generate 2D images used as input to CNN. This study used LeNet model (LeCun et al., 1998), with the two convolutional layers and 2 by 2 sub-sampling layers, to predict crop yields, and the experiments were conducted to obtain optimal parameters.

Stacked-Sparse AutoEncoder (SSAE): The structure of an autoencoder is similar to ANN, but the size of the input layer is the same as the size of the output layer. Since the hidden layer size is smaller than these, the feature information is compressed and extracted. Even if there are many hidden layer neurons, it can be learned by activating only some neurons randomly and sparsely (Ng., 2011). That means the 'sparse' term. 'Stacked' means that multiple sparse autoencoder are stacked, which is the same idea in which high-order feature information can be obtained by having multiple hidden layers in deep learning method. In this study, as with the previous neural network models, experiments were conducted to find the optimal parameters.

Long-Short Term Memory (LSTM): LSTM (Hochreiter & Schmidhuber., 1997) is a type of Recurrent Neural Networks (RNN) that has a cyclic structure in the hidden cell for processing sequential data, such as time series and language. The LSTM is designed to solve the vanishing gradient problem, which is a problem of deterioration in the learning ability of the RNN structure with long time lag data (Gers., 2002). In this study, experiments were performed to obtain the optimal parameters, and the activation function was set to hard sigmoid, and the hidden cell size, epoch, learning rate and batch size were adjusted to find optimal parameters.

Decision Tree: In this study, DT was used as a baseline model for comparison with other neural network models. Decision tree is a technique that is actively used in the field of machine learning, and it has also been widely used in studies to estimate crop yields (Johnson., 2014, Everingham et al., 2016). This study used Classification and Regression Tree (CART) algorithm with mean square error based node splitting criteria (Xu., 2005). In addition, to calculate the prediction accuracy of all the above mentioned models including decision tree, this study used N-folds cross validation method to compare average accuracy by dividing fold by year.

3. RESULT AND DISCUSSION

The model comparison is made using root mean square error of each model. Table 1 shows the prediction error of corn and soybean yield for five machine learning models for each year. Average RMSE shows that CNN has the lowest error, followed closely by SSAE. The average RMSE of LSTM model was worse than that of decision tree, the baseline model. Compared by year, LSTM model was found to be significantly inaccurate in 2005, 2010 and 2012. In the two years except 2012 that the error was large in all models, but only the error of LSTM model had the error lower than the average. Also, it was shown that the accuracy of all models in year 2012 was significantly inaccurate.

Table 1. Root mean square error results of yield prediction for five machine learning models.

Year	Corn yield prediction (RMSE %)					Soybean yield prediction (RMSE %)				
	DT	ANN	CNN	SSAE	LSTM	DT	ANN	CNN	SSAE	LSTM
2003	11.78	10.10	8.43	12.48	14.35	32.10	20.96	34.00	17.94	44.32
2004	9.16	8.37	6.79	6.11	7.00	11.62	10.82	9.96	8.56	14.46
2005	10.22	11.88	7.05	9.72	19.03	9.94	9.37	10.54	9.33	8.25
2006	10.88	8.84	7.58	8.56	15.56	11.20	7.51	6.91	8.54	9.34
2007	13.26	11.94	7.05	7.80	10.40	12.35	11.80	12.84	8.83	13.43
2008	9.10	7.65	7.53	7.80	9.82	14.26	11.57	7.53	10.40	18.60
2009	7.98	7.22	7.24	9.42	10.45	15.41	16.71	8.37	13.52	17.13
2010	16.85	10.66	10.13	10.70	23.47	11.12	9.87	8.73	7.95	12.38
2011	11.18	9.90	7.49	8.57	15.21	12.23	10.72	9.78	9.71	14.15
2012	26.79	24.49	24.97	22.03	28.94	16.33	15.74	12.91	9.93	11.09
2013	10.34	9.07	7.21	8.25	8.09	14.83	10.74	7.92	8.45	10.77

2014	11.28	6.20	8.18	8.07	6.35	14.90	11.27	8.46	10.06	8.53
2015	9.49	10.83	8.91	9.75	6.10	13.32	11.13	16.01	9.35	8.15
2016	11.50	8.73	12.43	9.07	5.88	13.23	9.89	11.62	8.19	8.26
Average	12.13	10.42	9.36	9.88	12.90	14.49	12.01	11.83	10.05	14.20

Unlike the corn yield prediction result, SSAE was the model with the lowest average RMSE. CNN and ANN model showed similar levels of accuracy, with the lowest error after SSAE. In addition, LSTM showed better average accuracy in the soybean yield prediction than the decision tree. In soybean yield prediction, year 2003 was a year where errors were significantly large in all models.

As mentioned above, this study also compared the yield prediction accuracies by periodic scenarios. Figure 1 shows average RMSE of corn and soybean yield prediction for six periodic scenarios. In Figure 1-(a), corn yield prediction, comparing the average RMSE of the four models among the six periodic scenarios, the best prediction result was obtained using data from May to September. In the same way, for soybean yield prediction, as shown in Figure 1-(b), prediction result from May to August showed the best average RMSE.

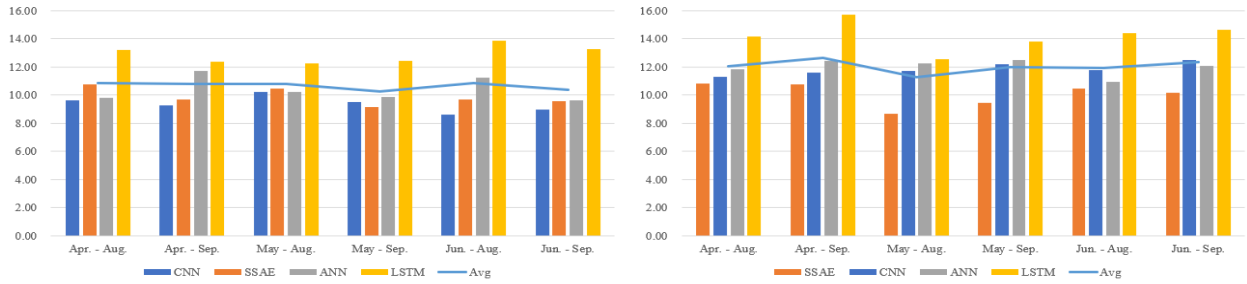

Figure 1. Average RMSE (%) result of four models by periodic scenarios; (a) Corn yield prediction, (b) Soybean yield prediction

The effects of natural disasters and pests also hinder the accuracy of crop yield prediction models. Corn yield prediction had a big error in 2012 and soybean in 2003. It could be assumed that unusual events, such as floods or droughts at the time of crop growth and harvest, result in very inaccurate predictions for a certain year. For example, in the case of corn, there was a big drought in the Midwest of the US in 2012 (Wan et al., 2015) and corn yield is more susceptible to drought in the Central US than soybean (Lobell et al., 2014). Therefore, it can be said that the reason for the inaccurate prediction of corn yield in 2012 was because of drought.

In corn yield scenario, predictions using the data up to September were more accurate than those obtained using the data up to August. Contrarily, for soybean, the input data period until August showed better prediction than the data including September. This can be explained by the study that the correlation between NDVI and daytime surface temperature and soybean falls earlier (mid-September) than corn (Johnson., 2014). So, it might be interpreted that data in September had a bad influence on the soybean yield prediction.

4. CONCLUSION

This study was conducted to predict crop yields for corn and soybean in Illinois and Iowa. ANN, CNN, SSAE, and LSTM were used to predict crop yield, and DT was used for the baseline prediction model. Four kinds of vegetation indexes from MODIS, weather data and yield statistics of each crop from 2003 to 2016 was used as input data and reference data. Also, data was divided into six periodic scenarios as input variables for each prediction model.

As a result, CNN model was most accurate over all with the lowest average %RMSE errors in corn (9.36%) and SSAE model in soybean (10.05%). The best periodic scenario for corn yield prediction was May to September, and for soybean was June to August. In addition, CNN, SSAE, and ANN showed significantly better prediction accuracy than the baseline model. Thus, neural network models seem to be yield better results in predicting crop yield than simple model. However, LSTM model did not predict better than the baseline model. This can be seen that LSTM is the model that performs prediction based on time series features, unlike other models, and thus shows different results.

However, this study has limitations in predicting the yields affected by unusual factors, such as abnormal weather conditions and crop diseases. Also, this study did not use data, such as soil properties, that are known to have a significant impact on crop growth and yield. In future work, a comparative study will be conducted involving more data and machine learning techniques.

ACKNOWLEDGMENTS

This work was carried out with the support of “Cooperative Research Program for Agriculture Science &

Technology Development (Project No. PJ009978)" Rural Development Administration, Republic of Korea.

REFERENCES

Chen, P., & Jing, Q., 2017. A comparison of two adaptive multivariate analysis methods (PLSR and ANN) for winter wheat yield forecasting using Landsat-8 OLI images. Advances in space research, 59(4), pp. 987-995.

Chlingaryan, A., Sukkarieh, S., & Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, pp. 61-69.

Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G., 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for sustainable development, 36(2), pp. 27.

Fernandes, J. L., Ebecken, N. F. F., & Esquerdo, J. C. D. M., 2017. Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. International journal of remote sensing, 38(16), pp. 4631-4644.

Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K., 2016. Rice crop yield prediction in India using support vector machines. In 13th International Joint Conference on Computer Science and Software Engineering, pp. 1-5.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J., 2002. Learning precise timing with LSTM recurrent networks. Journal of machine learning research, 3, pp. 115-143.

González Sánchez, A., Frausto Solís, J., & Ojeda Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction. Spanish Journal of Agricultural Research, 12(2), pp. 313-328.

Hochreiter, S., & Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp. 1735-1780.

Johnson, D. M., 2014. An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sensing of Environment, 141, pp. 116-128.

Kaul, M., Hill, R. L., & Walthall, C., 2005. Artificial neural networks for corn and soybean yield prediction. Agricultural Systems, 85(1), pp. 1-18.

Kuwata, K., & Shibasaki, R., 2016. Estimating corn yield in the United States with MODIS EVI and machine learning methods. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences, Vol.3(8), pp. 131-136.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp. 2278-2324.

Liu, J., Goering, C. E., & Tian, L., 2001. A neural network for setting target corn yields. Transactions of the ASAE, 44(3), pp. 705.

Lobell, D. B., Roberts, M. J., Schlenker, W., Braun, N., Little, B. B., Rejesus, R. M., & Hammer, G. L., 2014. Greater sensitivity to drought accompanies maize yield increase in the US Midwest. Science, 344(6183), pp. 516-519.

NCDC, 2019, National Climatic Data Center – Data Access, Retrieved September 2, 2019, from https://www.ncdc.noaa.gov/data-access.

Ng, A., 2011. Sparse autoencoder. CS294A Lecture notes, 72(2011), pp.1-19.

Ma, J. W., Nguyen, C. H., Lee, K., & Heo, J., 2016. Convolutional neural networks for rice yield estimation using MODIS and weather data: A case study for South Korea. Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, 34(5), pp. 525-534.

Ma, J.W., Nguyen, C.H., Lee, K. and Heo, J., 2018. Regional-scale rice-yield estimation using stacked auto-encoder with climatic and MODIS data: a case study of South Korea. International Journal of Remote Sensing, pp.1-21.

USDA NASS, 2019, United States Department of Agriculture - National Agricultural Statistics Service, Retrieved September 2, 2019, from https://quickstats.nass.usda.gov/.

USGS MODIS, 2019, MODIS data, Retrieved September 2, 2019, from https://modis.gsfc.nasa.gov/data/.

Wan, J., Qu, M., Hao, X., Motha, R., & Qu, J. J., 2015. Assessing the Impact of Year 2012 Drought on Corn Yield in the US Corn Belt Using Precipitation Data. Journal of Earth Science and Engineering, 5, pp. 333-337.

Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K., 2005. Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment, 97(3), pp. 322-336.

You, J., Li, X., Low, M., Lobell, D., & Ermon, S., 2017. Deep Gaussian process for crop yield prediction based on remote sensing data. In Thirty-First AAAI Conference on Artificial Intelligence, pp. 4559-4565.