Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain - August 19 - 21, 2024
Paper No. MVML 109
DOI: 10.11159/mvml24.109

GNN Approach for Maize Yield Prediction

Stefan Hačko[1], Đorđe Milanović[2], Milica Brkić[1], Sanja Brdar[1]
[1]BioSense Institute
Dr Zorana Đinđića 1, Novi Sad, Serbia
stefan.hacko@biosense.rs, milica.brkic@biosense.rs, sanja.brdar@biosens.rs
[2]IT HAPPENS LLC
Petra Drapšina 36, Novi Sad, Serbia
djordje.milanovic@ithappens.rs

Extended Abstract

As a result of rapid climate changes, combined with its wide application in various industries, the demand for maize is predicted to increase drastically in the coming decades. Developing new maize hybrids annually to enhance yield for various weather conditions and soil types presents an essential, yet exceptionally challenging endeavor for crop breeders. The process is hindered by insufficient data due to experimental costs, time constraints, and the limited number of locations available for hybrid planting. This information scarcity complicates critical decisions such as selecting top-performing hybrids for further experiments, registration, and commercialization.

This research aims to evaluate the performance of various maize hybrids in untested locations by utilizing historical yield, soil, and weather data. In the literature different approaches for fusing these three types of data can be found, ranging from the simple concatenation of features to using more complex matrix fusion algorithms [1, 2]. In this research, we formulate the yield prediction problem as an edge regression problem on a heterogeneous graph. A graph consisting of hybrid and environment nodes, where the weight of the edge connecting the two types of nodes represents the yield, is used to represent the data.

We used the well-known Syngenta Crop Challenge 2019 data. The data holds two datasets:

- The performance data - consisting of the hybrid ID, the environment ID (yearly location identifier, year, latitude, longitude, yield, planting date, harvest date, irrigation quantification, and 8 soil properties, all provided from 2008 to 2017.
- The weather data set - consisting of the following weather data: day length, precipitation, solar radiation, snow water equivalent, maximum temperature, minimum temperature and vapor pressure for all 1,560 environments daily.

In this research, we utilize a novel hybrid graph neural network (GNN) + XGBoost approach to make predictions on the aforementioned graph.
First, we train an encoder-decoder GNN, where the encoder layer consists of two graph convolutional layers (Graph Convolution layer followed by a SAGE convolution layer) [3], and the decoder is a shallow fully connected layer. Since we are working with a relatively small dataset, the shallow decoder's performance was not up to par with other techniques. To improve the performance, we encode the data using the trained GNN's encoder before training an XGBoost model [4] on the encoded data.

The models were trained and tested both on a per year basis, as well as on the whole dataset at once. With this approach, we observed an average RMSE improvement of 4% (ranging from 2.34% up to 8.267%) as compared to the current state-of-the-art approach.
It should be noted that even a slight increase in the model's performance combined with different agricultural portfolio optimization techniques can result in a non-negligible increase in the yield and profit for large-scale crop breeders. Given

the results of this research, we believe that graph representation of the relevant data will be key to ensuring the long-term sustainability and stability of crop breeding.

References

[1] M. Brkic, Stefan Hacko, Milos Radnovanovic, Vladimir Crnojevic and Sanja Brdar, Maize hybrids performance evaluation with Data Fusion by Matrix Factorization algorithm, *Engineering Applications of Artificial Intelligence*, Submitted 2024.

[2] S. Khaki and L. Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621, 2019.

[3] W. Hamilton, R.Ying, J. Leskovec, Inductive Representation Learning on Large Graphs, *Proceedings of the 31st International Conference on Neural Information Processing Systems,* 2017

[4] T. Chen and C.Guestring, XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016

[5] O. Marko, Sanja Brdar, Marko Panic, Predrag Lugonja and Vladimir Crnojevic Soybean varieties portfolio optimization based on yield prediction. *Computers and Electronics in Agriculture,* 127:467-474, 2016.