#014

Leveraging machine learning and environmental data to enhance genomic prediction in canola

Sakaria Liban [1, 3]

Lennard Ehrig [2]
Iulian Gabur [2]
Stephen Fox [3]
Evan Gillis [3]
Lon Rach [3]
Janice Duguid [3]
Jesse Mutcheson [3]
Mike Domaratzki [4]
Rod Snowdon [2]
Rob Duncan [1]

[1] University of Manitoba, Winnipeg, Canada

[2] University of Giessen, Giessen, Germany

[3] DL Seeds Inc., Morden, Canada

[4] Western University, London, Canada

Background:

Hybrid breeding programs have traditionally relied on per se parental nursery data for test cross selections despite the low association to hybrid performance for complex traits. The declining cost of genotyping technologies over the past decade has allowed for the collection of substantial genotypic data sets to complement phenotypic data thereby paving the way for large scale genomic selection models and Machine learning (ML) approaches.

Objective:

This study examines the predictive power of artificial neural networks and traditional genomic selection models to predict hybrid performance in canola (*Brassica napus*). Models include Genomic BLUP (gBLUP), Bayesian A/B/C/Lasso/Ridge Regression, Reproducing Kernel Hilbert space (RKHS), and Multilayer Perceptron (MLP) and Convoluted (CNN) Neural Networks. Traits with varying complexities were evaluated including days to flower, days to maturity, oil content, protein content, glucosinolate content, and yield.

Methods:

Phenotypic data for 5104 hybrids was collected from 2016 to 2020, representing 135 site-years across western Canada. Genotypic data was collected for parental genotypes using a 19K SNP array. Imputed environmental data was acquired for each site-year GPS coordinate using openweather.org for rain, snow, day length, humidity, temperature, wind, and calculated growing degree days. Hourly data was summarized and then normalized to seeding date. Additional growth window specific groups were added, totalling 7K environmental parameters per site-year. Mixed model analysis was used to generate means data for each hybrid, and for each site-year x hybrid for the environmental models. Genotype SNP data was imputed by linkage disequilibrium for a matrix of 19K SNPs x ~5104 hybrid phenotypic data dependent on trait. The GxE model consisted of a 26K (SNP + Environmental) x ~30K (site-year x phenotype) matrix varying with trait. Models were run in either R or python.

Results:

The variation in predictive accuracies was greater between traits with different complexities than between genomic prediction models. For genotype models, variation ranged from Pearson correlations values of 0.54 to 0.79 across traits. Overall neural network models provided an improvement in predictive accuracy on average but were more sensitive to size of training data. When environmental data was incorporated, we observed double digit gains in our GxE analysis with significant improvements on average across traits. Random subsampling for hybrid number demonstrated diminishing returns in model accuracy when using more than 1500 hybrids.

Conclusions:

While quantity and quality of phenotypic data remains more important than model selection, fine-tuned machine learning models provide an opportunity for incremental improvement in selection outcomes. The incorporation of environmental data provides further opportunity for improved genomic selection outcomes when data is segmented to biologically relevant growth windows. Aside from measurement error, we expect selection accuracy is limited by the genotypic variance captured by our SNP array and unmeasured environmental and management variables. In addition to the prediction of untested hybrid genotypes, GxE models allow for the prediction of optimal genotypes in untested environmental conditions, and insights to breeding for climate change or adapting to new environments.