FarmView: Regression Analysis of 2016 Sorghum Composition

Ben Parr (bparr@cmu.edu)

Machine Learning Department, Carnegie Mellon University

Introduction

Objective: Predict sorghum composition from accession, field location, ground robot, aerial drone, and harvest phenotype features.

- **Carnegie Mellon University**: Created a ground robot with multispectral cameras that traverses the rows of sorghum.
- **Near Earth Autonomy**: Created an aerial drone that is able to measure macro-scale field growth.
- **Clemson University**: Cultivated a field of sorghum in the clay soil of Pendleton, South Carolina. Biologists conducted composition experiments on the harvested sorghum.

Sorghum bicolor

- A preferred bioenergy candidate because it is drought-tolerant and highly productive.
- Has a diverse gene pool containing over 40,000 genetic varieties.
- Bioenergy sorghum breeders expect to double yield in the next five years alone.
- Production has grown by 66% in the last 50 years.

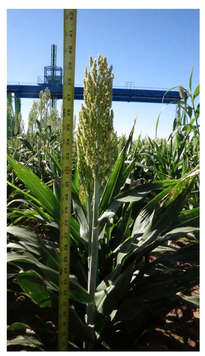

Figure 1: Photograph of growing sorghum, showing panicle, leaves and stem.

Simpson Farm, Pendleton, South Carolina

- Field partitioned into 5 meter by 1 meter subplots each containing genetic siblings.
- Composition results for 698 subplots (samples).

Figure 2: Overhead photograph of 2016 sorghum field, taken by the aerial drone.

Methods

The relatively small dataset allowed for trying multiple regressors (e.g. Random Forest, SVM, Nearest Neighbors, etc.). For each regressor:

- 10-fold cross validation.
- Missing values replace with -1.
- String values mapped to one-hot vector's (DictVectorizer).

Random Forests

- First formally defined by Breiman in 2001.
- Ensemble method which makes predictions by averaging the outputs of its constructed regression decision trees.
- Fast and perform well in practice (Fernández-Delgado et al. (2014), and Caruana et al. (2008))

Missingness Augmentation

The dataset has a significant amount of missing input values (more than half of the samples have at least 7 out of 29 input features missing).

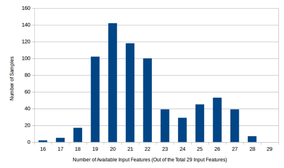

Figure 3: Missingness in the dataset. The dataset has no missing output values. So all missing values are from the 29 input features.

So, before training, augment the training set with training samples that have selected values removed and replaced with missing values so there are more training samples with missingness identical to missingness in the test set.

Example: Consider a dataset with three input values, and where X_train only contains three samples with values [7, 8, 9], [$MISSING$, 2, 3] and [4, 5, $MISSING$]. Augmenting the training sample with value [7, 8, 9] would result in:

[7, 8, 9] with $sample\_weight = \frac{4}{6}$
[$MISSING$, 8, 9] with $sample\_weight = \frac{1}{6}$
[7, 8, $MISSING$] with $sample\_weight = \frac{1}{6}$

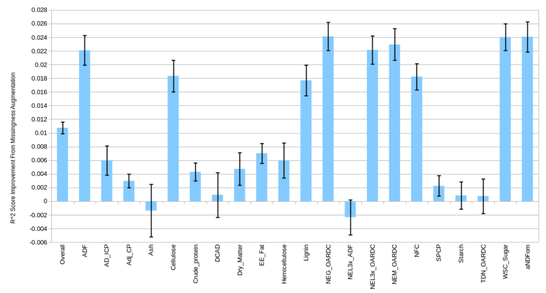

Figure 4: The difference between the mean r^2 scores with augmentation and mean r^2 scores without augmentation (100 trials each). So greater than 0.0 is an improvement from augmentation. The figure also includes the 99% confidence interval.

Results

Random forests performed best with an Overall r^2 score of 0.436 (includes 2.5% increase from missingness augmentation).

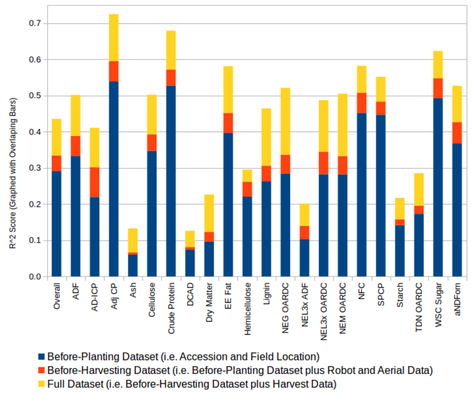

Figure 5: Results for each composition feature, and Overall. Each overlapping bar represents three separate dataset views.

Discussion

Upcoming sorghum field should include improvements from:

- Ground robot hyperspectral data of leaves.
- Aerial maps after resolving GPS issue with plot segmentation.

Conclusion

Missingness augmentation gave a 2.5% increase in Overall r^2 score with a minimal computational cost. Final Overall r^2 score is 0.436.

Project files: https://github.com/bparr/dap.

Finally, thank you to Dr. Artur Dubrawski for advising me on this project, Saswati Ray and Simon Heath for their project feedback, and the U.S. Department of Energy for funding FarmView.