Seed Stocking Via Multi-Task Learning and Two-Step Allocation

Yunhe Feng and Wenjun Zhou

University of Tennessee Knoxville
Finalist of the *Syngenta Crop Challenge 2017*

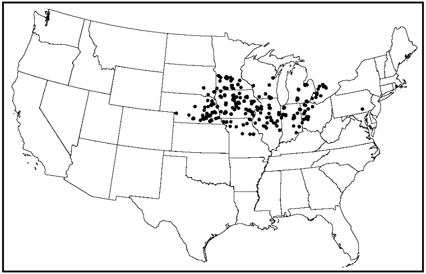
05/05/2017

Outline

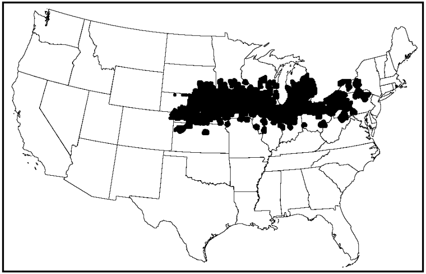
1. Problem Description and Overview
2. Why Multi-task Learning
3. How We Did It
   - Multi-task Learning Models
   - Prediction and Risk Analysis
   - Planning
4. Results and Evaluations
5. Conclusions & Future Work

Problem Description and Overview

- Problem: Which soybean seed variety (or mix of up to five varieties) should be stocked to meet the needs of farmers in the region?
- Two datasets: Experiment and Region



(a) Experiment



(b) Region

Figure 1: Locations in Datasets

Data Description

Table 1: Attributes in the Datasets

Category	Attributes	Meaning	Experiment	Region
Coordinates	Year	the year when the data are collected	2009–2015	2001–2015
	Lat. & Lon.	geo-coordinates of farmlands	583 locations	6490 locations
Weather	Temperature	sum of the daily temperatures	varies by year and location	
	Precipitation	sum of the daily precipitation		
	Solar Radiation	sum of the daily solar radiation		
Soil	CEC	Cation Exchange Capacity (cmol kg-1)	varies by location only; does not change by year	
	pH	log of H+ concentration in the soil		
	Organic Matter	the percentage of organic matter in soil		
	Soil Clay	the percentages of soil small particles		
	Silt	the percentages of soil medium particles		
	Sand	the percentages of soil large particles		
	PI*	the degree of suitability for growing crops		
Planting	Variety	seed variety to be evaluated	174 varieties	TBD
	Planting Date*	what day when the variety was planted	May-Sep.	
	Yield	crop productivity		

* has missing values

A High-Level Problem Analysis

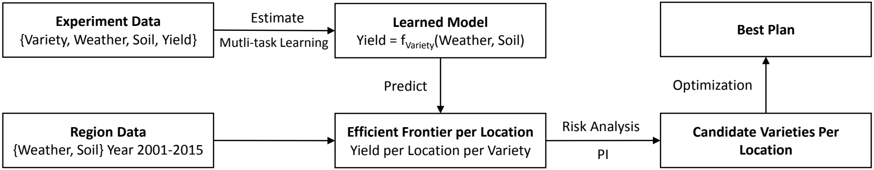


Figure 2: Framework Overview

- Estimate: multi-task learning (MTL)
- Prediction & Planning: modern portfolio theory (MPT)

Notations and Problem Formulation

- Suppose that the growing conditions are a p-dimensional vector (including soil conditions, weather conditions, etc.)
- Suppose that for variety i, n_i experiments were done in the past, each reported a yield y_{ij}, $j = 1, 2, \ldots, n_i$.
    - Then the training data matrix for this variety, X_i, would have a size $n_i \times (p + 1)$. The columns are p growing condition variables plus a constant.
    - Accordingly, the target/response data Y_i is a vector of length n_i.
- Our goal is to find a function $f_i : X_i \to Y_i$, so that we may estimate the yield of variety i under any growing condition.

Outline

1. Problem Description and Overview
2. Why Multi-task Learning
3. How We Did It
   - Multi-task Learning Models
   - Prediction and Risk Analysis
   - Planning
4. Results and Evaluations
5. Conclusions & Future Work

Why Multi-task Learning

The number of observations per seed variety in the “Experiment” dataset is highly skewed.

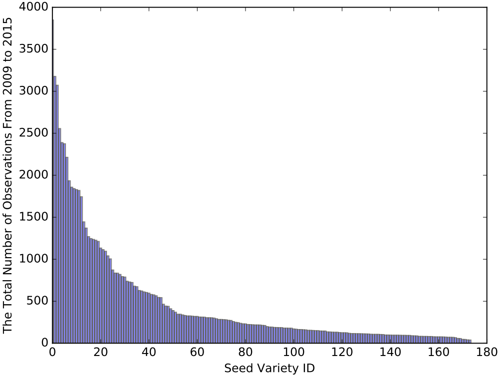


Figure 3: Number of Observations per Seed Variety

* More than 75% of the varieties have less than 500 observations.

Why Multi-task Learning

- By exploiting commonalities and differences across tasks, multi-task learning can improve learning efficiency and prediction accuracy for the task-specific models, when compared to training the models separately (Baxter, 2000; Thrun, 1996; Caruana, 1998).
- The commonalities are captured while doing the induction of multiple tasks, while the uniqueness is represented by the individual parameter for each task.
- Multi-task learning has demonstrated high performance in the scenarios where the number of labeled observations are limited because of high generation costs, such as in public health and bioinformatics studies (Xu et al., 2011; Zhang et al., 2012).

Outline

1. Problem Description and Overview
2. Why Multi-task Learning
3. How We Did It
   - Multi-task Learning Models
   - Prediction and Risk Analysis
   - Planning
4. Results and Evaluations
5. Conclusions & Future Work

Applying Multi-task Learning

- We treat the planting of each seed variety as an individual task, which takes multiple variables including weather and soil conditions as the input and the yield as output.
- We believe all these seed varieties share some latent commonalities because they all belong to the same crop category, which means they share the common ancestors and have a high level of crop-specific genetic similarity.
- Different seed varieties have their own specific minor characteristics, such as, strong drought resistance and salt tolerance.

Multi-task Learning Models

Mean-Regularized Multi-Task Learning

Assuming that all tasks (i.e., the planting of soybean varieties) are close to the average, and may deviate from the mean task.

Graph Based Multi-Task Learning

The closeness between each pair of tasks is represented by an affinity graph, which may be constructed by adding an edge between two tasks that are related (unweighted version), or by using the similarity score between each pair of tasks (weighted version).

Mean-Regularized Multi-Task Learning

Then, the mean-regularized multi-task learning formulation can be expressed as follows:

$$(1) \qquad \min_{W} \quad \sum_{i=1}^{V} \left\| X_i \vec{\beta}_i - Y_i \right\|_F^2 + \lambda \sum_{i=1}^{V} \left\| \vec{\beta}_i - \tfrac{1}{V} \textstyle\sum_{j=1}^{V} \vec{\beta}_j \right\|_1$$

where $\| \cdot \|_F^2$ represents the squared Frobenius norm, $\| \cdot \|_1$ represents the L_1-norm, λ is a regularizing parameter for mean-regularization, and V is the total number of varieties possible.

Our goal is to find the best $W = \left[\vec{\beta}_1, \vec{\beta}_2, \ldots, \vec{\beta}_V \right]$, where $\vec{\beta}_i$ is a coefficient vector of length $(p+1)$, which represents the (linear) model of task i.

Graph Based Multi-Task Learning

Suppose that G is the graph structure incorporating the similarities among the V tasks, the graph based multi-task learning formulation can be expressed as follows:

$$\min_{W} \sum_{i=1}^{V} \left\| X_i \vec{\beta}_i - Y_i \right\|_F^2 + \lambda_1 \left\| WG \right\|_F^2 + \lambda_2 \left\| W \right\|_1 \quad (2)$$

where λ_1 and λ_2 are regularization parameters.

To build the graph structure matrix G in this study, we first run a multi-task lasso with least squares loss (Tibshirani, 1996) to estimate the correlation coefficients. If the correlation coefficient of two tasks is larger than a threshold, an edge is added to connect the two tasks.

Performance Prediction Per Location

- Using the multi-task learned predictive model, we may predict the performance of each seed variety under any soil and weather conditions.
- Using the past 15-years' soil and weather data, we may estimate the yield for each year. Using these "data points," we could further estimate the overall performance (i.e., average yield) of each seed variety and risks (i.e., standard deviation of projected yield) under random weather conditions.
- With the mean and standard deviations, we may produce a scatterplot of all varieties. An example is provided in Figure 4.

Best Variety Composition Per Location

To find the best variety mix by striking a balance between yield and risk, we employ the asset allocation methodology based on modern portfolio theory (MPT).

- Determining the efficient frontier
- Identifying the maximum allowable risk

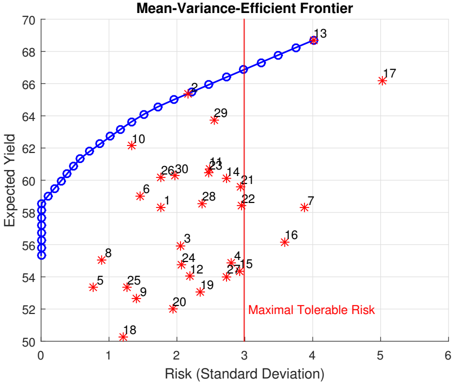


Figure 4: MPT Efficient Frontier

Determining Maximum Allowable Risk

- We do not know the risk preferences of farmers. However, we know that farmers' risk preferences may be inferred from the production index (PI) variable (Schaetzl et al., 2012).
- PI is the productivity index based upon soil classification. The greater PI value the greater the suitability for growing crops.
- Without knowing more details about the risk profiles, we will choose the risk level for each PI level using a linear scheme.

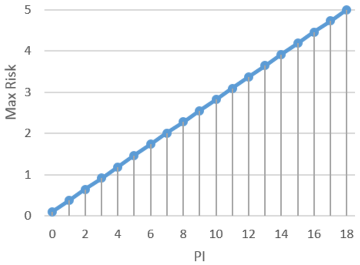


Figure 5: Maximum Allowable Risk

- Suppose that the maximal tolerable risk for a location with $PI = 0$ is R_0, and the maximal tolerable risk for a location with $PI = 18$ is R_{18}.

$$R_k = R_{18} - \frac{(18-k)(R_{18} - R_0)}{18}$$

- We selected $R_0 = 0.1$ and R_{18} is the risk value corresponding to the maximum possible return on the efficient frontier (MathWorks, 2017).

Seed Variety Allocation at One Location

Formally, suppose that if we only plant variety i next year, the yield will be $\mathcal{Y}_{li}$. For a location l with $PI = k$, then we optimize the following problem:

$$
\begin{aligned}
& \max_{\vec{w}_l} \quad \mathcal{Y}(\vec{w}_l) \equiv \vec{w}_l^T \mathcal{Y}_l, \\
(3) \qquad & \text{s.t.} \quad R^2(\vec{w}_l) \equiv \vec{w}_l^T Cov(\mathcal{Y}_l)\vec{w}_l \quad \leq \quad R_k^2, \\
& \qquad 0 \leq w_{li} \leq 1, \quad \forall i = 1, 2, \ldots, V; \sum_{i=1}^{V} w_{li} = 1.
\end{aligned}
$$

where $\mathcal{Y}_l = (\mathcal{Y}_{l1}, \mathcal{Y}_{l2}, \ldots, \mathcal{Y}_{lV})^T$ is a vector of length V that consists of the (random) yield of each variety at location l, and $\vec{w}_l = (w_{l1}, w_{l2}, \ldots, w_{lV})^T$ is a vector of length V that corresponds to the proportion of each variety in the optimal mix at location l. $\mathcal{Y}_l$ and $Cov(\mathcal{Y}_l)$ are estimated using our model predicted yield values for all varieties given location l's soil conditions and past 15 years' weather data.

Weight Optimization with Top K Seed Varieties

Aggregating the optimal seed varieties per growing location, we can determine the top K varieties.

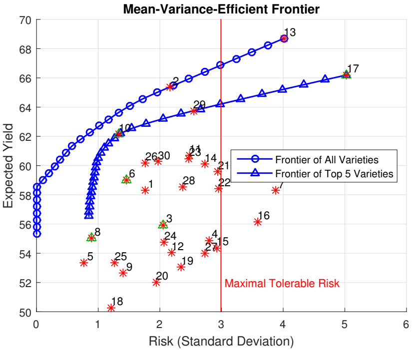


Figure 6: Objective Function Definition

- Using the top K varieties, we re-balance the weights and produce a new "efficient frontier".
- Take the intersection of vertical line $x = R_k$ and the new "efficient frontier" as the best variety combination for each growing location.
- Re-aggregating the demands of the top K varieties per growing region, we make the final decision.

Outline

1. Problem Description and Overview
2. Why Multi-task Learning
3. How We Did It
   - Multi-task Learning Models
   - Prediction and Risk Analysis
   - Planning
4. Results and Evaluations
5. Conclusions & Future Work

Performance and Benchmark

- Weighted RMSE (wRMSE) is commonly used in the multi-task regression.

$$(4) \qquad wRMSE = \frac{\sum_{i=1}^{t} \sqrt{\sum_{j=1}^{n_i} (X_{i,j} * W_i - Y_{i,j})^2 * n_i}}{\sum_{i=1}^{t} n_i}$$

- However, for any growing conditions in 'Region' datasets, the probability of being planted for each seed variety is same, i.e., a variety should not be given a larger probability of being planted just because it has a larger number of observations. So we propose the average RMSE (aRMSE) as follows:

$$(5) \qquad aRMSE = \frac{\sum_{i=1}^{t} \sqrt{\frac{\sum_{j=1}^{n_i} (X_{i,j} * W_i - Y_{i,j})^2}{n_i}}}{t}$$

Five-fold Cross Validation Results

Baselines: linear regression, lasso regression, ridge regression, stepwise regression, and SVM regression.

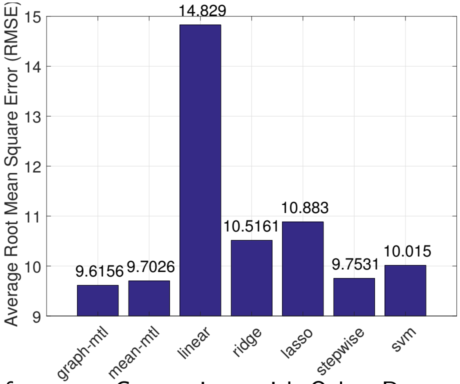


Figure 7: Performance Comparison with Other Regression Methods

RMSE of Individual Variety

For most of the varieties with a number of observations less than 500, GMTL outperforms linear regression. Recall that more than 75% of varieties have a number of observation less than 500.

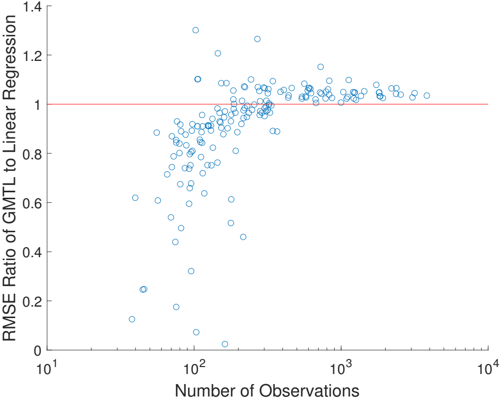


Figure 8: The RMSE Ratio of GMTL to Linear Regression

Top Varieties

Applying learned models on each growing location in the "Region" dataset, the total demand for each variety is visualized in Figure 9.

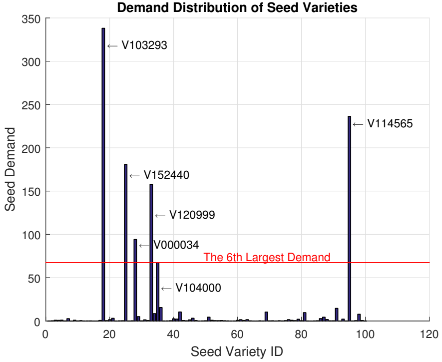


Figure 9: Seed Demand of Different Varieties

We can see that the top four (or top six) seed varieties by demand exceed other varieties by a large margin.

Seed Stocking Recommendations

After optimizing seed mix among top seeds for each location separately, and then aggregating all the 5 seed varieties together, we report the following aggregated proportions as in Figure 10.

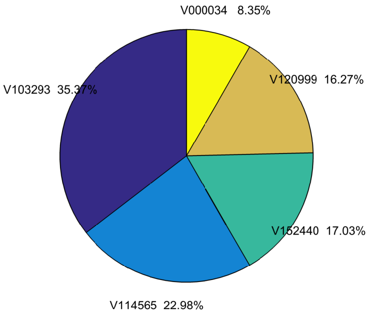


Figure 10: Proportion of Seed Varieties

- V000034 was dropped since its percentage is below 10%
- After redistributing the remaining four varieties, we recommend:
  - V103293: 36%
  - V114565: 23%
  - V152440: 24%
  - V120999: 17%

Outline

1. Problem Description and Overview
2. Why Multi-task Learning
3. How We Did It
   - Multi-task Learning Models
   - Prediction and Risk Analysis
   - Planning
4. Results and Evaluations
5. Conclusions & Future Work

Conclusions

- We used multi-task learning techniques to estimate yields of different locations by leveraging the commonalities as well as the uniqueness among different seed varieties.
- We determined the best mix of seeds for each location by seeking a tradeoff between yield and risk.
- We picked the top five varieties based on the aggregated best mix of individual locations.
- We re-balanced the yield and risk for each location by only growing a mix or a single of the top five varieties.

Further Improvements

- Gather more realistic estimates of maximum tolerable risks.
- Estimate weather variable distributions rather than using just the past 15 years' data points.
- Fine tuning parameters in a smaller granularity for the multi-task models.
- More evaluation baselines, such as regression trees and random forest.
- Other multi-task learning variations, such as robust MTL, relaxed ASO, and fused sparse group Lasso.
- More features, such as irrigation and fertilizer.

References

Baxter, J. (2000). A model of inductive bias learning. *J. Artif. Intell. Res.(JAIR)*, 12(3):149–198.

Caruana, R. (1998). Multitask learning. *Learning to learn*, pages 95–133.

MathWorks (2017). Arguments of numports in the 'portopt' function. `https://www.mathworks.com/help/finance/portopt.html?searchHighlight=portopt&s_tid=doc_srchtitle`.

Schaetzl, R. J., Krist, F. J., and Miller, B. A. (2012). A taxonomically based ordinal estimate of soil productivity for landscape-scale analyses. *Soil Science*, 177(4):288–299.

Thrun, S. (1996). Is learning the n-th thing any easier than learning the first? *Advances in neural information processing systems*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Xu, Q., Pan, S. J., Xue, H. H., and Yang, Q. (2011). Multitask learning for protein subcellular location prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(3):748–759.

Zhang, D., Shen, D., Initiative, A. D. N., et al. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage*, 59(2):895–907.