Literature Surveys for Capstone Project Phase-2 Review-1

Project Title: PerfectCrop - The Right Crop for your Soil
Project ID: PW22RBA01
Project Guide: Prof. Raghu B A Rao
Project Team: PES1201800051 Srish Srinivasan
PES1201800089 Akash Kumar Rao
PES1201800102 Vishruth Reddy
PES1201800291 Ishan Agarwal

1. Crop Yield Prediction using Deep Neural Networks

The paper throws light on the usage of deep neural networks to predict the crop yield using various parameters like the environmental conditions and the genotype of the crop. A root mean squared error of 12% was achieved which was reduced to 11% when accurate weather data was taken into account. The DNN model implemented gave better accuracy scores than the other models they tried to implement such as Lasso Regression, Regression Trees and SHallow Neural Networks. The main understanding from the paper was that environmental conditions (soil and atmospheric parameters) have more impact on the crop yield than the genotype. The crop under consideration was Maize.

The neural network they used had the following features:

- 21 hidden layers in each NN
- 50 neurons in each layer
- 3,00,000 maximum iterations
- Batch normalization and Adam Optimizer were used

The above hyperparameters reduced overfitting and improved the accuracy of the model.

Genotype, soil and weather parameters were compared to see the extent of influence. The following table provides the details:

Model	Training RMSE	Training correlation coefficient (%)	Validation RMSE	Validation correlation coefficient (%)
DNN(G)	21.74	20.26	21.72	15.09
DNN(S)	15.28	73.37	15.49	72.04
DNN(W)	14.26	76.98	14.96	72.60
Average	24.40	0.0	23.14	0.0

Here, DNN(x) refers to the performance of the models on individual parameters without the consideration of the other parameters.

From this, we can see that:

- Soil and weather DNNs have similar performance measures
- DNN(S) and DNN(W) have higher accuracies than DNN of genotype.

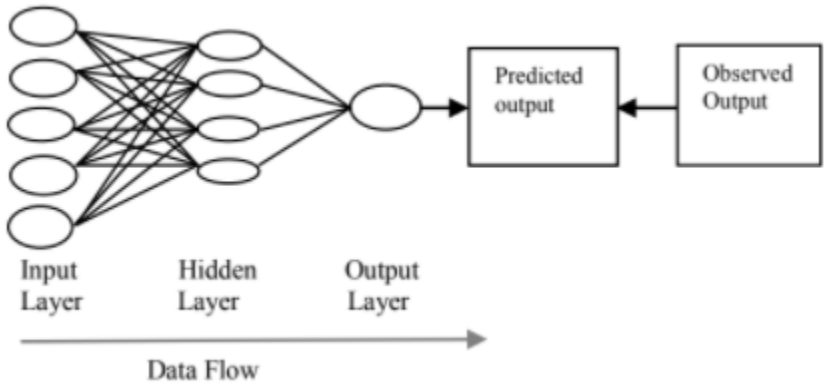
We can conclude that weather and soil parameters are essential and can't be excluded as their impact on the yield has great significance. The black box property was eliminated by performing feature selection on the trained model making use of back propagation methodology.

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

2. Agriculture Crop Yield Prediction using Artificial Neural Network Approach

The paper provides insights upon the need for prediction. Growing crops takes months of hard work and tremendous amounts of resources. If the right crop can be recommended, it will be very beneficial. Fuzzy systems, Genetic algorithms and Artificial neural Networks have provided great efficiency and accuracy in prediction. Reducing the scope of study to only ANN, the computational model they used was feed forward back propagation neural network.

The reason for choosing ANN was because it can find the factor that is most impacting the yield. Just like any other ANN, they have used a model that contains an input layer, the interconnected hidden layer and an output layer as shown below:



Using the back propagation algorithm, which uses the gradient descent method, they were able to achieve a high accuracy.
They were able to predict cotton, sugarcane, jowar, bajra, soybean, corn, wheat, rice and groundnut using ANN with a good accuracy score.

The following are the requirements for the crops to grow:

Crop	PH	Nitrogen (ppm)	Depth (ppm)	Temp (°C)	Rainfall (cm)
Cotton	7-8.5	40	30	27-33	700-1200
Sugarcane	6.5-7.5	40	60	20-50	750-1200
Jawar	6.0-8.5	132-180	50-20	25-30	800-1000
Bajara	7-8.5	120	15	28-32	400-750
Soyabean	6.5-7.5	37	15-20	25-33	700-1000
Corn	7.5-8.5	60-120	5	13-30	500-600
Wheat	6-8.5	80-150	15-20	16-22	25-180
Rice	5.5-8.5	50	50-20	22-25	1000-1500
Groundnut	6-7.5	25	20	24-27	500-1250

Citation:
Dahikar, S.S. and Rode, S.V., 2014. Agricultural crop yield prediction using artificial neural network approach. *International journal of innovative research in electrical, electronics, instrumentation and control engineering*, *2*(1), pp.683-686.

3. Influence of Minority Class Instance Types on SMOTE Imbalanced Data Oversampling

The author of this paper aims to understand the impact of instances belonging to the minority class on the performance of SMOTE by performing a selective preprocessing of these instances. Most contemporary classification techniques make an assumption that all the classes in the training dataset have almost equal number of instances under their name. However, this assumption need not always come true and in the case of an imbalance, the classifier will end up making predictions in the favour of the majority class. So, our goal must be to ensure that the classifier performs well when it comes to the minority classes and at the same time make sure that their performance with the majority classes doesn't drop by a large margin.
The data-level solution to this problem aims at balancing the data that is used to train the model. The algorithm-level solution to this problem looks into the specifics of the algorithm being used and if a drawback or 2 are found, it tries to improve the algorithm in those areas. This problem is also being tackled by the use of an ensemble of classifiers which could put up a better show collectively as compared to working as individual entities.

Data-level approaches work independent of the classifier being used. Under this category, SMOTE was a very popular algorithm that was being employed despite the fact it too had some drawbacks. One of them is to treat all minority instances equally. But being more selective when it comes to oversampling the minority instances yielded a better overall accuracy accordingly. Fortunately, there are few solutions that focus on this drawback of SMOTE. Borderline-SMOTE focused on those instances that sit close to the class demarcations. This idea of Borderline-SMOTE was further developed into a technique called ADASYN which could dynamically pick those instances that are more problematic to the classifier. Safe-Level-SMOTE allocates weights to instances based on how much the majority class influences these instances and makes use of these weights for the generation of artificial instances. A technique called SPIDER focuses on those instances that overlap with the majority class.

Algorithm 1: Synthetic Minority Over-Sampling Technique aka "SMOTE"

Function *SMOTE* ($D_{minority}$, $N_{percent}$, k)

$D_{smoted} \longleftarrow [\,]$

for $i \leftarrow 1$ **to** $nrow(D_{minority})$ **do**

$nn \longleftarrow kNN(D_i, D_{minority}, k)$

$N_i \longleftarrow \lfloor N_{percent}/100 \rfloor$

while N_i != 0 **do**

$neighbour \longleftarrow select\_random(nn)$

$gap \longleftarrow range\_random(0, 1)$

$diff \longleftarrow neighbour$ - D_i

$synth \longleftarrow D_i + gap * diff$

$D_{smoted} \longleftarrow append(D_{smoted}, synth)$

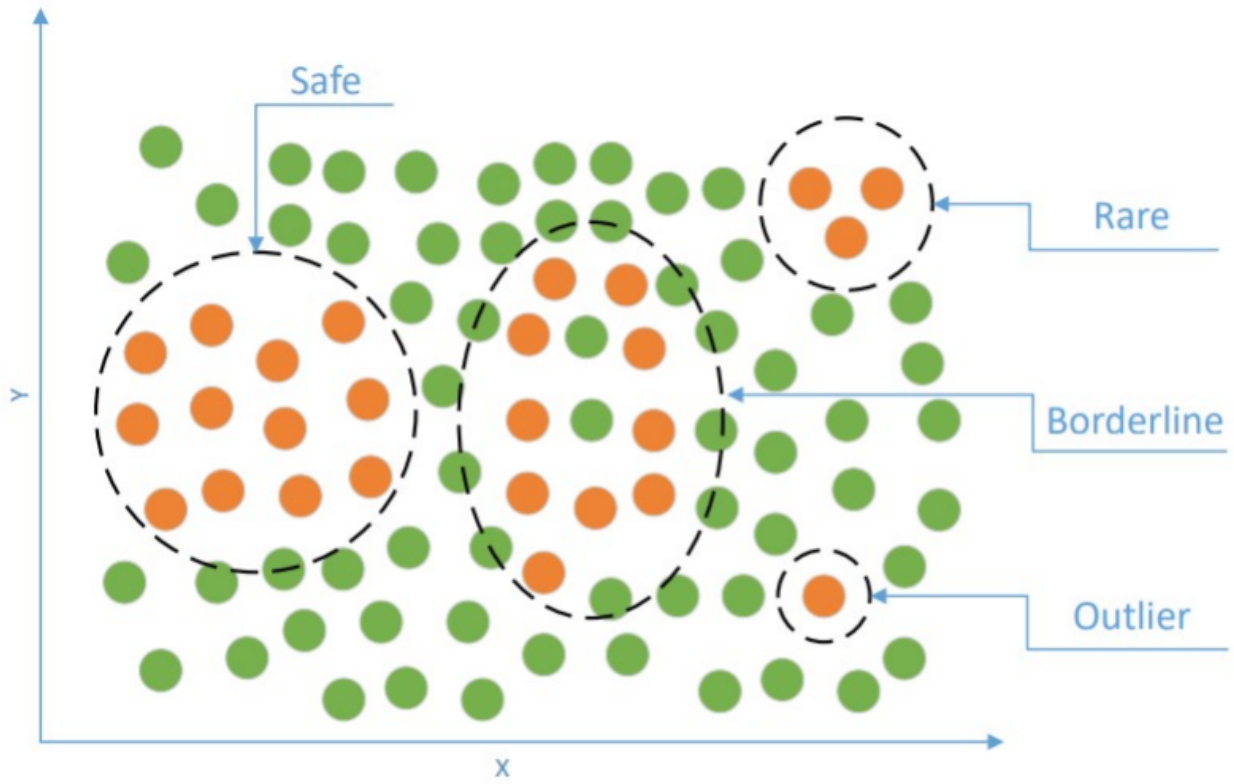
$N_i \longleftarrow N_i$ - 1

end

end

return $\longleftarrow D_{smoted}$

SMOTE is computationally quite complex and its memory requirements are quite significant. When working with imbalanced data, apart from the class disparity, the characteristics of the minority class instances are also imperative. In most cases, these minority class instances have a crystal clear demarcation through the formation of heterogeneous structures. Some of these structure types are as follows.



The algorithm used to obtain the above segregation is as follows and the distance metric used here is the Heterogenous Value Difference Metric and not Euclidean distance.

Algorithm 2: Neighbourhood analysis for determining types of minority class instances

Function *Types (D, k)*

```
types ⟵ [ ]
D_minority ⟵ get_minority(D)
D_majority ⟵ get_majority(D)
foreach x_i in D_minority do
    neighbours ⟵ kNN_HVDM(x_i, D, k)
    N_minority ⟵ minority_samples(neighbours)
    if N_minority ⩾ ⌊0.8k⌋ then
        types_i ⟵ "safe"
    end
    else if N_minority ⩾ ⌊0.5k⌋ then
        types_i ⟵ "borderline"
    end
    else if N_minority ⩾ ⌊0.2k⌋ then
        types_i ⟵ "rare"
    end
    else
        types_i ⟵ "outlier"
    end
end
return ⟵ types
```

After having determined the types of minority class instances, a base classifier was be tested in different preprocessing configurations which is achieved by oversampling only selected types of minority class instances. In conclusion, the author stated that data-driven oversampling improved the performance of the classifier as compared to a uniform oversampling technique.

Therefore, the imbalance ratio is not the only cause for learning obstacles, but the properties of the instances belonging to the minority classes need to be looked into as well and thereby enabling SMOTE to oversample only selected instances from the dataset.

Citation:
Skryjomski, P. and Krawczyk, B., 2017, October. Influence of minority class instance types on SMOTE imbalanced data oversampling. In first international workshop on learning with imbalanced domains: theory and applications (pp. 7-21). PMLR.

4. Yield Prediction with Machine Learning Algorithms and Satellite Images

The author has used given satellite images over Iran in this paper and used it along with some simple regression models to obtain the yield prediction over a particular region of Iran along with the optimal time. The author has chosen a specific crop i.e. barley and has predicted the yield for the same.
The author has used the remote sensing and climate data from the Google Earth Engine (GEE) platform these were integrated with four machine learning algorithms i.e., backpropagation neural network (BPNN), decision tree (DT), gaussian process regression (GPR) and K-nearest neighbour regression (KNN) algorithms.
It is further seen that the two indexes used by the author are NDVI and EVI which are calculated as follows:

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

$$EVI = 2.5 * \frac{(NIR - RED)}{(NIR + C_1 * RED - C_2 * BLUE + L)}$$

Compared to NDVI, EVI does not saturate at high canopy densities, can reduce canopy background signal and atmospheric influence, and improve vegetation dynamics in high biomass areas. A combination of NDVI and EVI can provide more information about the crop yield, which helps yield prediction.
Considering three evaluation indicators (r2,RMSE and MAE), GPR, DT, BPNN and KNN models showed higher accuracy, with r2 (0.84–0.69) and RMSE (< 737 kg ha −1) and MAE (< 650 kg ha −1), respectively. Therefore, these models are very suitable for predicting atmospheric performance in southern Iran.
Results proved the GPR model to be the best in terms of accuracy and generalization compared to the other models. The results showed that the accuracy depended on the location and also time. It was seen that the GPR model could accurately predict barley crops a month before harvesting time. It was also seen that EVI, temperature and precipitation played the most important part in the model.

Citation:
Sharifi, A., 2020. Yield prediction with machine learning algorithms and satellite images. Journal of the Science of Food and Agriculture.

5. Crop yield estimation using Satellite Images: Comparison of Linear and Non-linear Models

The goal of this research was to develop and evaluate linear and non-linear models to estimate crop yield from satellite data. The authors used images from the Landsat and SPOT satellites to obtain soybean and corn yield in the central region of Córdoba (Argentina). This was also compared with field data collected over the places to check the validity of the model.
Since in the cultivated regions where the chosen lands are small the data from Landsat and SPOT satellites were sufficient due to their spatial resolution. Four images, two from each, were examined which were taken on days with clear weather and after the required image processing ,the subset of all the said were taken in order to cover the entire area. This was further used for training and testing the data on various ML models. MLR models were used where the yield was calculated as:

$$\text{Yield} = \sum_{i=1}^{k} a_i v_i + b$$

where the regression variables are:
vi = surface reflectance of i band of SPOT/Landsat satellite and model constants are ai and b.

Neural networks were also used with an input layer whose number of neurons was equal to the number of bands considered in each model; a hidden layer was designed with the same number of neurons as the respective input layer. The output layer was built with only one neuron that indicates the calculated crop yield.
It was found on comparing all of these that all regression and neural network models developed to estimate yield provided a good fit with measured yield.
This further proves that satellite images can be used to accurately determine yield of a crop and also neural networks provide better accuracy than machine learning models in most cases .
Hence, The possibility of combining satellite images with climatologic or soil data to improve the performance of yield estimation, is the next step to be explored.

Citation:
Sayago, S. and Bocco, M., 2018. Crop yield estimation using satellite images: comparison of linear and non-linear models. AgriScientia, 35(1), pp.1-9.

6. Understanding Satellite-Imagery-Based Crop Yield Predictions

In this paper the authors have reviewed previous work done over the USA county and have tried on improving an existing model. In the previous model the authors had effectively done dimensionality reduction and then the result of their Central Neural Network was fed in a Deep Gaussian process to eliminate changes due to spatial correlation of yields between counties. The authors successfully are able to get better results by adding more layers to the CNN and hence simplify the process by eliminating the need for a Deep Gaussian process.
The authors used the data from the MODIs satellites available on the Google Earth Engine. Seven spectral and two temperature bands(for night and day) were taken.They had also used ground yield data available for the crops in question. Since an important aspect of this was to deal with spatial correlation the crops chosen were soybean and corn as they are grown around similar regions.
The further made the data ready for use by using permutation invariance which basically allows the particular pixel's value to be the important factor instead of location so that two pixels of same value would have same weightage regardless of the location thereby meaning that adjacent farms or no farms in surrounding would have no effect on them. This was converted into a histogram matrix containing the necessary features and discarding the non essential ones ; this histogram was given as input to their model. The prediction was done using:

$$\tilde{y}_{corn} = \frac{\hat{y}_{soy} - \hat{\mu}_{soy}}{\hat{\sigma}_{soy}} \cdot \hat{\sigma}_{corn} + \hat{\mu}_{corn}$$

$$\tilde{y}_{soy} = \frac{\hat{y}_{corn} - \hat{\mu}_{corn}}{\hat{\sigma}_{corn}} \cdot \hat{\sigma}_{soy} + \hat{\mu}_{soy}$$

To predict corn yield they first standardized the distribution of soybean yield predictions made using the soybean model to zero mean and unit variance using the mean and standard deviation of soybean yield in the validation set and then rescaling the distribution to the mean and variance of corn yield in the validation set. They got better accuracy for their model and unlike previous work they used more than one crop.
The conclusion was that prediction accuracy increases with complex models and it was also proved that pixels could be used to differentiate between the different crops.

Citation:
Sabini, M., Rusak, G. and Ross, B., 2017. Understanding Satellite-Imagery-Based Crop Yield Predictions. Stanford.

7. Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques

In this paper the authors have used various different vegetation indices which are normally used in remote sensing. It is seen that healthy crops have a characteristic of strong red band

absorption and reflectance of the near-infrared band(NIR) ,this property is taken into consideration in the different indices.
The indices taken into consideration are:

- NDVI (normalized difference vegetation index):
  This index is made up of the red and NIR band. This is seen to be effective in large areas rather than small research areas (which happens to be the case here)

$$NDVI = \frac{\rho_{ir} - \rho_r}{\rho_{ir} + \rho_r}$$

- GVI (green vegetation index):
  This is based on the green and near-infrared band(NIR). This index according to previous work surveyed could account for soil water content and varying soil background.

$$GVI = \frac{\rho_{ir} - \rho_g}{\rho_{ir} + \rho_g}$$

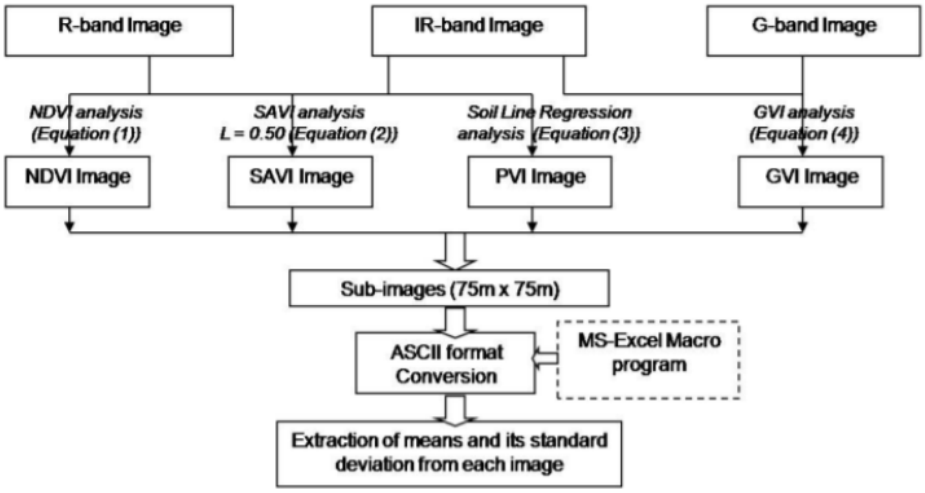
- PVI (perpendicular vegetation index):
  This helps in correction of soil reflectance more than NDVI

$$PVI = \sqrt{(\rho G_{ir,s} - \rho P_{ir})^2 + (\rho G_{r,s} \vdash \rho P_r)^2}$$

- SAVI (soil adjusted vegetation index):
  This like PVI is a distance based index and corrects soil brightness .This came to be after a soil calibration factor being included in this and is an index midway between NDVI and PVI.

$$SAVI = \left[\frac{(\rho_{ir} - \rho_r)}{(\rho_{ir} + \rho_r + L)} \right] x \ (1 + L)$$

The chosen area was in North Dakota USA in the Oaks Irrigation Test Area.
The input data were the different indexes obtained from above mentioned formulae which were separately then passed into different models.
This is as follows:



These images were then passed through (some data transformation done in some cases). This study showed that PVI showed the most accuracy in this case and also improved on the others by data transformation compared to other indices but it is also attributed to it being over a small area.

Citation:

Panda, S.S., Ames, D.P. and Panigrahi, S., 2010. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Remote Sensing, 2(3), pp.673-696.