# Crop Data Analytics using Image and Non-image Features

February 1, 2017

## 1 Background

Agriculture has been the key sector in all major economies and especially most important one for a country like India. Over a past couple of decades, technology has been revolutionized and much of the focus was shifted on to the technological advancement without much contribution to the agricultural domain. With the increasing population, the need for an increased amount of crop production can be observed. The annual food production of India was around 50 million tons in 1950-51 and touched 240 million tons by 2010-11. The demand for food grains is expected to reach 280 million tons by 2020. Various factors like newly emerging crop diseases, pests, untimely and excessive use of pesticides and fertilizers by farmers and extreme climatic conditions have made it difficult to meet the required needs. Considering all these problems, exploiting technological advancement to solve agriculture related problems seems to be of high importance. However, much of the technological advancement has not been translated successfully to agricultural domain. Nevertheless, there have been many successful attempts to deliver technology-enabled solutions to our farmers. During last decade, several research efforts were being made to build IT-based systems to analyze the agricultural information and provide necessary solutions for various problems that arise in agricultural domain. In 2004, an agricultural advisory system called eSagu was developed which aims at providing farm-specific agricultural expert advice to the farmers at regular. The esagu system requires several inputs including the images of the crop, weather details, soil conditions etc and by analyzing this information, the expert generates the advice. In this project, we are going to use the non-image data received as input to generate the required farm-specific advice.

## 2 Summary of the problem

This project proposes to analyze the data collected by eSagu. In particular, the idea is to analyze the non-image data coming from farms along with weather conditioning and attribute set by domain experts. Eventually, the target is to learn the mapping between visual and non-visual data by employing existing mathematical models like Neural Network (NN), probabilistic graphical model (PGM) and topic modeling for developing a prescriptive analytic solution in agriculture domain.

## 3 Related work

In India, there were about 120 million farm holdings by 2005 and the number has been growing year by year. Estimates indicated that 60 per cent of farmers do not access any source of information for advanced agricultural technologies resulting in huge adoption gap. During last decade, several organizations and government departments have built web portals and there were research efforts to build agriculture information delivery systems like aAQUA, e-Krishi, AGRISNET, Kisan Call Centres, and Digital Green. In 2004, one such research effort called "Esagu" was made at IIIT, Hyderabad.

In Esagu system, the agricultural experts receive latest information about the crop situation in the form of both image and non image data. Image data includes several pictures capturing several views of the farm and non image data includes details like the weather information, soil conditions, location of the farm etc. There will be a coordinator assigned to a specific region and he collects the above mentioned information from various farms and passes this information, along with observation form signed by him, to the agricultural expert. By analyzing the non image data, the agricultural expert identifies the problem, suggests a solution and passes the advice to a higher level expert, who in turn, verifies the solution, makes changes if necessary and finally approves the advice which is delivered to the farmer.

The idea of this project is to minimize the human effort in generating the required advice i.e, to automatize the process to the possible extent. The known information about the crops and various problems that may arise under specific conditions depending upon the crop species, season, locality of the farm, type of the soil, rainfall etc.. is used to predict the possible cause of the problem as soon as the information regarding the farm is received. If the database is unable to find a matching cause i.e, the information about the problem is not in our database yet, the database will be updated and this new cause will be added to the list. After each advice, feedback is collected from the respective farms and will be used to measure the accuracy of the system and modify it whenever required.

Whenever the data about a farm is provided to the esagu system, it suggests the most possible cause of the problem to the Agricultural expert who in turn verifies it and passes it on to the higher level expert.

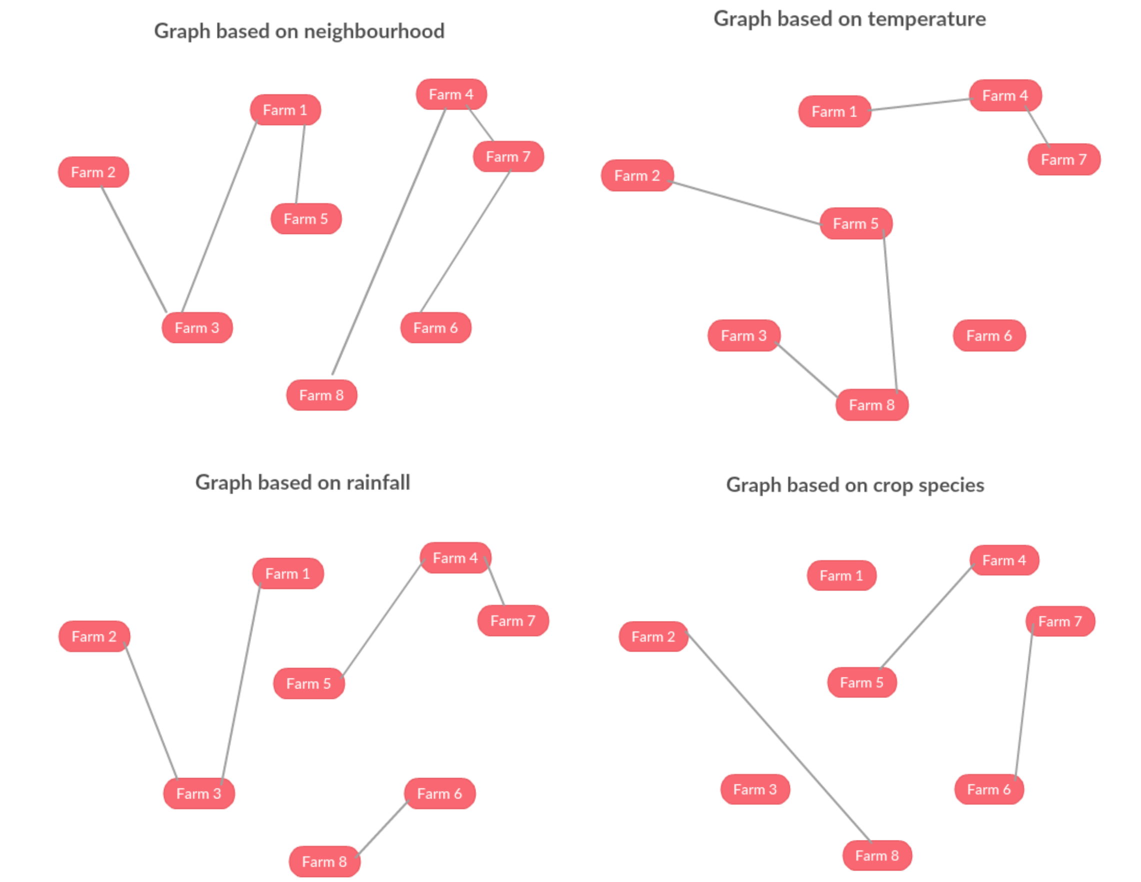
## 4 Expected contributions

Our idea of solving this problem will be done in a number of stages. Since our approach is using some machine learning algorithm, the first step is data preprocessing. Data preprocessing is further divided into a number of steps. Data cleaning involves filling in missing values, smoothing the noisy data and resolving the inconsistencies in the data. Next stage in data preprocessing is data integration i.e, data with different representations are put together and conflicts within the data are resolved. Another important step is data transformation. Here, data is normalized, aggregated and generalized. To store the data in a database, data reduction is very important. Finally, to make use of the data, discretization is done in which the number of values of a continuous attribute are reduced by dividing the range of attribute intervals.

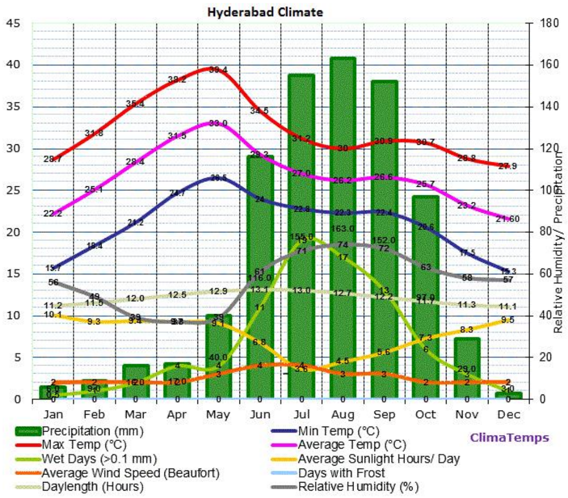
Now we have the processed data, we can use techniques like graph modeling to find the relations between various factors responsible for the growth of healthy crops. There are several approaches to design the required recommender system. One approach which might fit this problem is collecting and analyzing a large amount of information regarding the growth of various crops under various weather/soil conditions, at different stages of growth and the problems that may arise under certain conditions and predict how a specific crop will get affected based on the known data. For example, Diamondback moth is common pest affecting the Brinjal, Cabbage and Cauliflower plants. With this known information, if the collected information is regarding those plants, our focus would immediately turn towards that pest and check if its true. More and more information will be added to the database after every encounter with a new problem.

In this project, we are going to use a probabilistic graphical model(PGM) in which the nodes represent various farms in the specified domain and the edges between the nodes indicate the dependency between the farms based on various factors. This model shows the conditional dependency between the farms and since we have to consider various factors like the location of the farms, their neighbourhood, respective weather conditions, soil conditions, species of crops etc while modeling the system, initially we will get multiple graphs with the edges of each graph denoting one of the factors mentioned. The idea is to analyze all the dependencies from various graphs obtained and finally be able to find out the node which affects the specified node the most. There might be a strong dependency between two nodes when we consider one of the factors, which might not be observed if the remaining factors are considered and sometimes it is possible that no dependency can be observed between two nodes unless and until we take into account that one key factor which dominates the combined effect of remaining factors. Our model has to consider all these possible cases and we have to access multiple views of the data before arriving at the solution. And, at each node denoting the respective farm, we get several signals when we plot each of the mentioned factors against time which we can use to predict how a specific factor varies over the next period of time and how it might affect the crop. All these graphs and plots have to be analyzed once the input data is received.

### 4.1 An example of graph modeling




### 4.2 An example of graph signaling




We will use topic modeling to identify various key words, present in the input report, that correspond to various symptoms and find out whether any previously known problem matches with the identified symptoms. Using similar approach, we might be able to find the best matching farm for a given one in terms of various factors like average temperature, rainfall, type of soil etc. We can also use multiview learning (MVL) to train the model. The particular MVL technique depends on the type of data. If the dimension is very high, we should use dimensionality reduction techniques like canonical correlation analysis(CCA) and if the entire data is labeled, we will use supervised learning techniques. If we have both labeled and unlabeled data, we use semi-supervised learning.

Similarly, we can use other MVL techniques. By employing such machine learning techniques, our system will be able to predict the problem for a given farm.

## 5 Deliverables

- Data pre-processing
- Building the understanding of the NN, graphical and topic models
- Training NN and topic models
- Developing dashboard for interactive visualization of the data and prescriptive analytics service.

## 6 References


- http://insait.in/AIPA2012/articles/005.pdf
- http://insait.in/AIPA2012/articles/009.pdf
- http://www.saravananraj.net/wp-content/uploads/2014/12/45_AFITA_ICT-for-Agricultural-extensio India.pdf
- http://www.cst.ecnu.edu.cn/~slsun/pubs/MVML_final.pdf