Feature Selection using Random Forest Technique for the prediction of pest attack in cotton crops.

Shanmuga Priya.S
Department of CSE,
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
ss_priya@cb.amrita.edu

Abinaya.M
Department of CSE,
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
abinayamurugan94@gmail.com

Abstract— Incorporating the technology into the Agriculture field for triggering the growth and identifying the pest/disease in various crops is the state of art. Extreme Changes in the weather parameters during the growth phase of crops causes serious threats which tends to go in search for new adaptive measures. The main aim of this work is to predict the occurrence of the pest in cotton crop based on such weather factors using clustering techniques. A change in any of these meteorological factors fluctuates the infestation of the pest between higher and lower. The values of the weather parameters computed which cause the occurrence of these pests can be used for prediction and precautionary measures can be taken ahead.

Keywords— correlation; RFE; cotton; crop; pest; clustering

I. INTRODUCTION

Cotton is a major cash crop in India. It is necessary to protect the cotton plant from all types of attacks that hinder their growth so that it will be helpful in enhancing the economy. It is found that the pest attack for the cotton plant is mainly due to the environmental conditions. The environmental conditions directly or indirectly affect the plant. These conditions have to be identified well in advance.

The development of the pattern recognition techniques provides feasibility for predicting such pest occurrences with a validated input. If prediction is made, control measures can be taken to protect the crop and losses can be reduced in advance. This helps in minimizing the usage of pesticides and insecticides. The main techniques that are adopted in finding the pest occurrence are through the feature selection method and clustering. The incidence of the pest is categorized as high and low. It is known that categorization of the incidence is based on the features that contribute to the occurrence of pest. It is not necessary that every parameter should contribute as a whole part, for a certain cases only some parameters affect the results. In order to know the features significantly having a role in the occurrence of the pest, the feature selection algorithms are deployed. The Correlation, Random Forest are the feature selection algorithms implemented and k-means clustering is applied to predict at what range of values the pest will start affecting the Cotton plant.

II. RELATED WORKS

The prediction method was discussed in “Weather based prediction of pests in cotton crops”, by K.V.Raghavendra, D.S.Bhupal Naik, S.Venkatramanaphanikumar, S. Deva Kumar, S.V.Rama Krishna [2] which was submitted to the Sixth International Conference on Computational Intelligence and Communication Networks. The authors provide a detailed study on the incidence of pests influenced by the weather parameters for the period of 2006 to 2010 at Acharya N. G. Ranga Agricultural University, Lam farm.
Guntur. The pests chosen were Thrips, Aphids, Jassid and White fly. The pests were collected at weekly mean wise and on randomly selected 20 plants the incidence was taken. They are counted from top, middle and bottom of the plant.

The ANN method was discussed in “An Intelligent System For Predicting Thrips Tabaci Linde Pest Population Dynamics Allied To Cotton Crop”, by Jyothi patil , Dr A.Govardhan, Dr V.D.Mytri [3] tell us about the data mining techniques that are used in Agriculture field to predict the Occurrence of Thrips pest in Cotton Crop. This prediction is mainly focused on pest management programs to prevent the future occurrence by which amount of crop lost can be minimized. Various climatic condition affect the pest control measures when an unexpected event occurs. Pest data was taken from University of Agriculture Science, Raichur, India. Data reduction and data elimination techniques are applied on data as basic steps of pre-processing and the pre-processed data is normalized which is now good for training using neural network. Feed forward algorithm with Back-Propagation is applied to predict the occurrence. The results obtained using this method is able to predict the number of pest occurred.

The performance accuracy and runtime of clustering techniques is proposed in “Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time”, by Nidhi Singh,.Divakar Singh [14] which was submitted to the International Journal of Computer Science and Information Technologies. This paper discusses about two clustering techniques. It gives a comparative study about the performance and accuracy for k-means and hierarchical clustering. The data sets were collected from UCI repository of machine learning databases. The performance of hierarchical and k-means clustering has been calculated on WEKA data mining tool.

III. MATERIALS AND METHODS

The Proposed Methodology makes use of feature selection algorithms to identify the best response variables in terms of climatic parameters for identification of pest incidence. For the chosen response variables, clustering are deployed to find the pattern for low and high infestations.

Data Procurement

Cotton is a warm weather crop and sucking pests are found to be deleterious during early season of the cotton plant growth and development. The major sucking pests affecting the yield of cotton crops are identified to be Aphids[Aphis gossiji Glover], Thrips[Thrips tabaci Lindemann], Jassids[Amrasca biguttula], and Whiteflies[Bemisia tabaci]. These pests start occurring right from the sowing and if not monitored their population grow in numbers and ultimately affect the entire crop. Climatic conditions vary drastically with respect to the sowing and harvesting period, where sowing commences from Mar to May and harvesting is done in the month of Nov and Dec. AICRP has conducted an experimental setup to study the incidence of pests related to climatic conditions. Procurement of such pest data along with meteorological data for each sucking pests are collected for 5 years [1]. The collected data consists of weather parameters and its influence on the pest Thrips. The data in confined to one particular area in north zone Sri ganga nagar. The data is analyzed by using the analytical tool.

Feature Selection

There are many meteorological parameters contributing to the influence of weather. To identify the essential parameters affecting the pest, feature selection algorithms are deployed. In the current Scenario, the experimental results are based on Filter Selection methods. We have proposed a wrapper selection method called Recursive Feature Elimination (RFE) which construct disjoint decision tree to select the best feature. We have also compared the RFE method with the filter selection methods.

A. Filter Selection Methods

The filter selection method used here is Correlation.

The method works by the following steps:

1. A Matrix M[5][5] is constructed showing the correlation values between the features.
2. Assign M[i][j]=NULL ∀ i = 1 to 5 and i = j.
3. Compute mean of each column using the \( M = \frac{\sum_{i=1}^{5} M[i][j]}{4} \) ∀ i = 1 to 5  \[3.1\]
4. Reorder the matrix based on rank of the mean values M in descending order. Highest rank mean value is taken as correlation cutoff.
5. Check if the M[i][j]> cutoff, then
6. Repeat steps 5 until all correlation of features are examined.

B. Random Forest Approach

Random forest, a feature selection algorithm used for training the dataset because the dependency of the features is high. The method works by the following steps:

1. The algorithm is executed and the calculation of each variable importance based on the Mean decrease in Accuracy and Mean decrease in Node impurity is noted.
2. Ranking and scoring of the variables using the above two types are done.
3. Selecting the high score variables as the new variables.
4. Rerun the algorithm and compute their error rate.
Clustering

The significant parameters which are identified by the feature selection are clustered by using the clustering algorithm. The k means clustering algorithm is used for the clustering process to differentiate the incidence of occurrence of thrips as high and low.

1. The occurrence of pest is being differentiated as high and low hence the number of clusters is specified to be two.
2. The algorithm assigns the centroids randomly and to try with the different centroids the nstart attribute is specified.
3. To make the algorithm run for many iterations maximum iteration attribute is specified.
4. The distance measure used to calculate for the formation of cluster is Euclidian distance.
5. As the algorithm runs, the data points are re-assigned to the cluster whose value are closest to the centroids. Then the value for the new centroid is calculated.
6. The previous step is repeated until the cluster variation can be further reduced.

IV. RESULTS AND DISCUSSIONS

A. Correlation

With the correlation as the feature selection method it is not possible to select the important parameters because the cut off obtained by analysis for the dataset will not result in correct significant factors. It is less than the ideally used cut off for feature selection. Table.1 shows the Correlation result and its ranking.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Temp.</td>
<td>1.000</td>
<td>0.674</td>
<td>-0.844</td>
<td>-0.35</td>
<td>0.271</td>
</tr>
<tr>
<td>Min Temp.</td>
<td>0.674</td>
<td>1.000</td>
<td>-0.409</td>
<td>0.307</td>
<td>0.296</td>
</tr>
<tr>
<td>Morn Hum.</td>
<td>-0.844</td>
<td>-0.409</td>
<td>1.000</td>
<td>0.631</td>
<td>0.433</td>
</tr>
<tr>
<td>Even Hum.</td>
<td>-0.352</td>
<td>0.307</td>
<td>0.631</td>
<td>1.000</td>
<td>0.825</td>
</tr>
<tr>
<td>Rainfall</td>
<td>-0.271</td>
<td>0.296</td>
<td>0.434</td>
<td>0.825</td>
<td>1.000</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.19</td>
<td>0.217</td>
<td>-0.047</td>
<td>0.352</td>
<td>0.320</td>
</tr>
<tr>
<td>Rank</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Table.1 Correlation matrix

B. Random Forest

The parameters Maximum Temperature, Minimum Temperature, Morning Humidity, Evening Humidity and Rainfall are taken as V1,V2,V3,V4 and V5 respectively.

As per the result analysis the parameters Maximum Temperature (V1), Minimum temperature (V2) and rainfall (V5) are found to be the most significant attributes which leads to the infestation pests in cotton plants. Figure.3 illustrates the number of trees reducing the error rate. Figure.1-2 illustrates the parameters that are significant and the stability of the tree. Thus the random forest algorithm efficiently works in sighting out the important parameters required for the clustering process.

![Figure.1 Significant factors identification using Random Forest](image1)

![Figure.2 Number of trees Vs Error](image2)

![Figure.3 Maximum Temperature Vs Thrips](image3)
performed with the chosen parameters to have a prediction on the values where the infestation of the pest is higher. This can be further extended to apply Multiple Linear Regression techniques to include the time and severity of the pest, which can be made useful to set up an alarm system and take precautionary measures accordingly.

### References


[8] S.A.Ramesh Kumar, Dr.K.Ramesh Kumar, “A Study On Paddy Crops Disease Prediction Using Data Mining Techniques”.


