Pesticide recommendation system for cotton crop diseases due to the climatic changes

1. Mrs.’s. Pudumalar, 2. Ms.K.S. Suriya, 3. Mr. Ramanujam, 
4. Dr.S. Muthuramalingam

Department of Information Technology Thiagarajar College of Engineering Madurai, Tamil nadu, India

ABSTRACT

Data mining is a process of extracting knowledge from a vast database using tools and techniques. Data mining plays an important role in decision making on issues related to many real-time problems such as business, education, agriculture etc. Data mining in agriculture helps the farmers to decide on crop yield ratio, water resource management, pesticides management and fertilizer management. Nowadays, climatic change is one of the challenging problems in agriculture which has a greater impact on productivity. Many researchers have contributed in the field of agriculture data mining i) To predict crop productivity, ii) water management, iii) air pollution using the naïve bias and decision tree algorithms. The Proposed work is to predict the diseases due to Climatic changes and recommended pesticide for the disease. Decision tree algorithm is used to develop a recommendation system which helps to the farmer in the usage of pesticide for the incidence of crop diseases.

Keywords: Data mining, prediction algorithms, prediction.

Date of Submission: Date, 08 January 2018
Date of Accepted: 28 January 2018

I. INTRODUCTION

DATA mining is a process of extracting knowledge from a vast database using various data mining tools and techniques. Data mining plays a significant role in making decisions on issues related to real-time problems such as business, education, agriculture. While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the connection between the two processes. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. It consists of major five elements such as 1) Extract, transform, and load transaction data onto the data warehouse system. 2) Store and manage the data in a multidimensional database system. 3) Provide data access to information technology professional and business analysts. 4) Analyze the data by application software. 5)

Present the data in a useful format, such as a graph or table. Predictive analysis is one of the important areas of the data mining. It is used to extract the information from data and use it to predict patterns and behavior. Now a day’s predictive analysis plays a major role in many fields like education, healthcare, social media, supermarket, online shopping and agriculture. All the fields have a massive amount of data, from which useful information is extracted. From the learned information, the future behavior and pattern will be predicted. This is broadly used to make decisions and helps to increase the profit percentage of the particular field. Major matrices of predictive analysis are accurate. Agriculture is the backbone of the India. In our country, productivity cost is less than investment cost. Nowadays, climatic change is one of the major challenging factor that farmer faces in the agriculture which has a great impact on productivity. Due to the climatic changes, many diseases affect the crop yield productivity. Climatic pests, bacteria, fungus, and virus are mainly affecting the crop growth. It will lead a huge loss for the farmer. If we predict the diseases due to the climatic change, it helps the farmer to improve the productivity and increases the profit. In India, cotton is the most important fiber crop, cultivated at about 11.73 Million Hectares; producing 39.00 Million Bales. In the year of 2014-2015, the yield was 565 Kg/Hectare. Many factors affect cotton production. Among them, the pest incidence and crop diseases caused by meteorological factor has a major impact. The meteorological factor includes maximum temperature, minimum temperature, rain fall, humidity, sunshine, evaporation. Due to changes in meteorological factors the cotton plants are affected by the pest. It mainly affects the leaves of the plant. The pest will attack the plant at five stages such as adult, egg, larva, and nymph. If we predict pest attack due to the climatic changes, it may help the farmer to protect crop against pests.
This is achieved by applying information technology in agriculture field. Especially, data mining techniques help in the agriculture field to predict climatic change and pest attack for the particular crop. So many researchers have worked in the field of agriculture data mining. i) To predict crop productivity, ii) water management, iii) air pollution. The productivity cost is less than cultivation cost which will affect the farmers. Using data mining tools and techniques, an efficient recommended system to rectify the issues in agriculture can be developed. The performance of the system is measured with metrics like accuracy, precision, space, time, and usefulness.

II. LITERATURE SURVEY

In [1] agriculture, a comparison of prediction models is made for short-term in rice stripe virus disease and its association with biological and meteorological factors. Rice stripe virus is a virus that affects the rice crops in lower and middle of the Yangtze River which leads to decrease in rice productivity. Data mining techniques are applied to predict the rice stripe virus disease associated biological and meteorological factors. In this paper different approach such as support vector machine, artificial neutral networks and regression were used for prediction. The accuracy of the above approaches are compared (Table 1). The average prediction accuracy of support vector machines, back propagation neural network and stepwise regression are 98.95\%, 93.75\% and 77.35\%. Mapping the [2] crop using survey data. The East African highlands banana data is used for mapping the bacterial wilt. This paper addresses the challenges of disease incidence mapping by applying a geo statistical methodology. The datasets are collected from framework of the USAID Crop Crisis Control Program, the case of bacterial wilt of banana (BWB) caused by the bacteria at the place of East African highlands. Regression analysis is applied to improve the spatial prediction of disease incidence in areas with low sampling density the target variables which were correlated in a logistic regression. It depends upon the environmental variables, for which maps were available. The accuracy of prediction is quantified using cross validation. This method generates high-resolution maps for the spatial distribution of infected crop, which is done by using local survey datasets and spatially in-depth ecological covariates.

In [6] the Smart beehive preliminary decision tree is used to monitor and analyze the honey bee health. This method is effective and suitable for all the weather conditions, but do not prevent from the beekeeper’s activities which are critical for hive maintenance. Honey bees are a major part of the food chain as the most important pollinator for a most of the crops. Protecting the population of honey bees worldwide, as well as enabling them to maximize their productivity, is an important concern. Dataset is collected using wireless network sensor. The dataset describes the internal conditions and colony activity of the bee accurately. The parameters such as CO2, pollutant gases, O2, relative humidity acceleration and temperature are monitored continuously as it has greater impact. Additional information on Weather condition such as sunshine, Rain, and temperature are also to lookup in different dimensions of analysis. Real time dataset is collected at a field-deployed beehive, a biological analysis was undertaken to classify ten important hive states. The decision tree is used to construct the model. Decision tree gives an accuracy of 95.58\% and finally the correlation is made between the meteorological factors and beehive data. This paper insight into the mechanism for autonomous maintenance of bee colonies. Data mining techniques are used to analyze crop yield prediction. Crop yield prediction is one of the challenges that every farmer meets. Data mining techniques such as multiple linear regressions and density based clustering techniques are used for prediction. Dataset collected between the periods of 1955-2009, Godavari district of Andhra Pradesh in India is used. The multi linear regression is applied to the dataset and the density based clustering is used to verify and analyze the result of the multiple linear regressions. The similar process was adopted for all the districts of Andhra Pradesh to improve and authenticate the legality of yield prediction, which are helpful for the farmers of Andhra Pradesh for the prediction of a specific crop.

The Climatic indicator [8] indicated the crop infection risk. The climate change impacts on five major foliar fungal diseases in Northern France. A decision support tool can help the stakeholder to anticipate future trends and the associated risks of crop diseases. Future agriculture planning emerges with assessment toll to evaluate fungal diseases pressure and crop regional suitability under projected future climatic change. Two climatic indicators AIE and NID to quantify the potential effects of weather on the intensity and occurrences of pathogen infection. Average Infection Efficiency (AIE) is monthly average of daily calculated infection efficiencies. Number of Infection Days (NID) quantifies the potential effects of weather on the occurrence of infection. Modified Weibull distribution is to develop the surface response function. Finally the impacts of climate change on the infection risk are assessed by calculating AIE and NID for 5 pathogens. The process based epidemiological models are applied in large scale with various pathogens. The [9] Prediction of climate
change impacts on cotton yields in Greece. The eight climatic models are used for the AquaCrop crop simulation model. The discriminant function is used for the analysis of the crop model. The Land and Water Division of FAO have developed the crop water productivity model which is called as Aqua Crop. It simulates yield response to water for herbaceous crops. It will mainly well-matched to address environment where water is a key limiting thing in crop production. The prediction of the climatic impacts on cotton yields, design a climatic model and apply the stepwise discriminant function analysis. The average cotton yields record in the seven locations between 1961 and 1990. The monthly averages of the precipitation and temperature for the growing period of cotton (April to November) in the seven study areas during the period 1961 to 1990 is which the data provided by Hellenic National Meteorological Service. By running the models C4I and DMI the relative impacts of climate change on seed cotton yield in the different areas were predicted. In the agriculture field, crop productivity cost is less than investment cost. In this paper [18] zero algorithms is used for the estimation of agriculture resource and pesticide using WEKA tool. The algorithm will extract the results from the target attribute. The dataset consist of cultivated area, crop yield, irrigation types, humidity in particular area, usage of pesticides, fertilizers and labor cost. The result helps farmer to reduce the chemical cost and also retain the fertility of the soil. This helps the farmer to decide usage of pesticide and fertilizer which will improve the profit for the farmer.

III. METHODOLOGY

The Proposed work is to predict the occurrence of crop diseases due to Climatic changes and recommend pesticide. Decision tree algorithm is used in decision-making process. The dataset is collected from cotton pest DSS and central institute for cotton research. The meteorological data are collected from the Cotton pest DSS. The dataset contains 17 attributes and 3017 instances. The attributes are crop name, location, pest population, maximum temperature, minimum temperature, rainfall, humidity, evaporation, sunshine value, and pests, stages of the attack, symptoms, recommended pesticide. The above attributes are used to predict the diseases and pesticide for the related diseases. The data preprocessing technique is applied to convert unstructured data into structured data. Preprocessing such as replacing the missing values, float to integer conversion are used to prepare the data for further processing. Feature selection is used to select the attributes that holds greater promise for prediction is done with alpha investing algorithm. The decision tree algorithm utilizes information gain, gain ratio and gini index for tree generation. Gain ratio is used to build the tree as it gives an efficient tree than gini index and information gain. GUI is designed for pesticide recommendation system. The system helps to the farmer to get information related to the pest attack due to sudden climatic changes, incidence of diseases and usage of amount and type of pesticide.

3.1 Architecture: Crop disease data is collected, preprocessed, and feature selected to drive the processed data from the raw data. Each process uses some techniques to form a processed data. Data collected from the real world which is in unstructured form. The unstructured data are converted into the structured data by using data preprocessing techniques. The processed data is given as input to the feature selection. Feature selection is used to select the promising attribute. Alpha investing algorithm is used for feature selection. Alpha investing algorithm reduces the false ratio of the dataset by setting threshold value to select the dataset. Each feature (attribute) will add one by one and the false ratio is measured when false ratio of an attribute increase, then the corresponding attribute is omitted from the dataset. The accuracy increases based on the feature selection technique. If we select the wrong attributes, the accuracy of the algorithm automatically decreases. After the feature selection process, the data is called as processed data and it is ready for analysis. Decision tree algorithm is used to classify the dataset. The decision tree generates rules and tree. The tree explains the traversal of the dataset and guides in decision making. The rules are induced and stored in the GUI. The user will submit his query on the GUI; checks for rules and predicts the pests at the particular weather condition. Based on pests, stage of the attack and symptoms and the pesticide is recommended for the corresponding pest. Fig 1 clearly explain the detailed design of the proposed architecture.

![Figure 1: architecture Diagram](image)

www.i jmrem.com  
IJMREM  
Page 27
3.2 Dataset Collection: The dataset is collected from cotton pest DSS and central institute for cotton research. The meteorological data are collected from the Cotton pest DSS. It holds data for many locations. Dataset collected from Akola and Coimbatore centres are considered for the cotton pest prediction. Four cotton pest meteorological values are used. Pest such as aphid, mealy bug, mirid bug, and Thrips are identified. Each pest will occur on the particular weather condition at every year. The dataset consists of 52 weeks, each week the pests will arrive at particular weather condition. All the week will be considered in the dataset. Dataset consists of crop, location, week, pest population, maximum temperature, minimum temperature, Relative humidity (1%), Relative humidity (2%), rainfall, sunshine value; wind speed, evaporation, and pests name. The stage of the attack, symptoms and pesticide are collected from the central institute for cotton research. In our study, pest, their stage of attack, symptoms, pesticide and pesticide level are given to the farmer.

3.3 Data Pre-processing: Data are the key ingredients in which every processing step has to be carried out carefully. The dataset enters into the KDD process. First, the data should be selected related to the application. To make the dataset efficient, data pre-processing is done. Since dataset may contain outliers and missing data. This may affect the accuracy of the prediction. Data preprocessing is the initial step. Next the cleaned data are transformed using data transformation techniques and then data analysis is done with data mining techniques. The KDD process helps us to extract knowledge using interpretation and evaluation. Visual representation of induced rules is done. The dataset contains 17 attributes and 3017 instances. The attributes are crop name, location, week, pest population, maximum temperature, minimum temperature, rainfall, humidity, evaporation, sunshine value, and pests, stages of the attack, symptoms, recommended pesticide and pesticide level.

3.3.1 Replace missing value: The numerical missing values are replaced by using mean values. Because each week the weather conditions are changed, one week it will be zero and another week it may have high value. So it is better to replace the missing values with mean. The String missing value should be replaced with unknown. Because replacing it with some other string values on the missing place, the accuracy may get affected. The pest population contains lots of missing values ie 90% of the values are missing and hence remove the pest population attributes in the dataset.

3.3.2 Float to integer: The dataset consists of meteorological attributes which contains the float values as JavaScript doesn’t support float values it is converted into integer values. The floating number is rounded off to its nearest integer.

3.3 Feature selection: Feature selection is used to select the best attributes and forms a subset of the dataset. This avoids unnecessary attributes in the dataset which will reduce the accuracy and efficiency of the system. Hence feature selection is done to improve the accuracy level. Alpha investing algorithm is used for feature selection. It works by setting a threshold value and scans each attribute sequentially. It measures the false ratio of the dataset. In this algorithm, the attributes are added one by one and corresponding threshold value is measured, when the threshold value increases, the attribute is removed from the dataset.

3.4 Prediction using decision tree: The data learning is used to represent the data in the hierarchical model which will help to know the root attribute and assign a weight to the attribute. The hierarchical model also priories the attributes in the dataset. The decision tree is used to learn and predict the pests for the dataset. A decision tree is suitable for the large dataset to clearly explain the dataset in the tree structure. The decision tree algorithm is generating rules for the prediction and tree. The data mining tool used for prediction is Rapid Miner. The dataset is loaded into the tool and it generates a large decision tree.

Fig 2 clearly shows the complete decision tree for the processed dataset. The gain ratio is used to build the decision tree. The extension of the information gain is called as gain ratio. Gain ratio defined as,

\[
\text{Gain ratio (A)} = \frac{\text{Gain (A)}}{\text{Split Info (A)}}
\]
The rules are induced from the decision tree. The root node of the decision tree is location. Based on the location and meteorological factor, the pest of attack is obtained. Some pests satisfy only three conditions. An user-friendly GUI has been developed to support the farmers. Fig 3 shows the graphical user interface for the farmers.

Pesticide recommendation system works as follows: 1. The user enters the location and weather condition

The System: shows the type of pest attacked. 3. The user selects the stage of the attack and the noticed symptoms. 4. The system recommends the pesticide and its level. Decision tree algorithm provides 85% accuracy.

IV. RESULT AND DISCUSSION

The dataset splits into two datasets. The first dataset consists of meteorological factors and pests names without location. It named as multiclass without location. And another dataset consists of pest name, stage of attack, symptoms, pesticide and level of pesticides. Integration of datasets is a challenging factor.
4.1 Comparison of accuracy between binary class and multiclass dataset: The dataset consists of multiclass values. These values are converted into binary class dataset and accuracy is measured. Binary class dataset gives 80% accuracy. The accuracy of each binary class dataset is measured and compared with multiclass dataset which will give 40.09% accuracy. Fig 4 clearly shows the accuracy of multi-class and the binary class. The accuracy measure for multiclass dataset without a location gives 40.09% accuracy. If we add the location attributes to the dataset, it will give 85% accuracy. Fig 5 shows the accuracy measure after feature selection. Fig 6 shows the accuracy of the multi-class dataset with and without feature selection.

**Figure 4:** Comparison of accuracy between binary and multi-class dataset before feature selection

**Figure 5:** Comparison of accuracy after feature selection

**Figure 6:** Comparison of accuracy with and without feature selection
4.2 comparison of accuracy between Decision tree and other data mining algorithms. The multiclass with location dataset are used to analyze the other algorithms’ performance and accuracy. We will consider decision stump, random tree, and random forest. The accuracy of the other algorithms are as follow, decision stump (66.67%), random tree (51.92%), random forest (77.31%), decision tree (85.51%). The accuracy compares to other algorithms, the decision tree will give high accuracy. Future work is to improve the decision tree accuracy. Fig 7 shows the Comparison of accuracy between decision tree and other data mining algorithms.

![Comparison of accuracy between decision tree and other data mining algorithms](image)

**Figure 7:** Comparison of accuracy between decision tree and other data mining algorithms

V. CONCLUSION

Data mining in agriculture helps to decide on crop yield production, water resource management, pesticides management and fertilizer management. Nowadays, climatic change is one of the challenging problems in the agriculture which has a great impact on productivity. Many researchers have worked in the field of agriculture data mining i) To predict crop productivity, ii) Water management, iii) Air pollution by using some data mining algorithms. The proposed work is to predict the diseases due to climatic changes and recommended fertilizer and pesticide for the disease and measuring the accuracy of the algorithm. Decision tree algorithm is used to create a model for the decision-making. Thus the decision making will help to the farmer to increase their food productivity and predict the profit and loss percentage for the future. So far so many problems are addressed using data mining techniques is discussed in the literature survey. The proposed work recommends pesticide for cotton crop diseases due to the climatic change using the decision tree. Rules are induced from decision tree and stored, with which the system recommends the pesticide. Initially, the cotton crop pest is predicted and corresponding recommendation is provided to the end user. The decision tree gives an accuracy of 85% for large dataset.

VI. FUTURE WORK

The proposed work considers only cotton crop with five pests, the future work can be extended to Rice crop with a large number of pests. The accuracy of the proposed system is 85.5% which can be improved by using decision tree with the genetic algorithm.

**REFERENCE**


