Development of Decision Tree Induction Model Using Sorghum Multi Location Data For Classification and Prediction

P. Mukesh[1], Sameen S.Fathima[2], Vasumathi, D[3], . Pratheepa [4] and Kalaisekar [1]
1 Directorate of Sorghum Research (DSR), Rajendranagar, Hyderabad -500 030
2 Professor and Head, College of Engineering, Osmania University, Hyderabad
3 Associate Professor, Additional controller of Examinations, College of Engg.,JNTU, Hyderabad
4 Scientist (Sr. Sc.), National Bureau of Agriculturally Important Insects, Bangalore 560 024, India

Abstract

Sorghum bicolor (L.) Moench is one of the major food grains under impoverished conditions in the semi arid tropics and is the fifth most important staple food crop for more than 500 million people in about 30 countries (***Kelley et al. 1993***) for food, feed, fodder and fuel. Sorghum is a C4 plant to be eco-friendly in role of climate change. To improve sorghum yield and quality there has been huge amount of multiplication data is generated every year in the sorghum research system which formulates thousands of data sets for analysis and predicts useful information and to make decisions. There is compilation of summarized annual data in the reports of All India Coordinated Sorghum Improvement Project (AICSIP) such are available make t in the form of hard copy for various important trails of sorghum. The data is not available in soft digitized to retrieve data for which an effort has been made to develop a data base program to retrieve database information on sorghum research and application of data mining technique.

The data-mining technique, for predicting grain yield, pest incidence etc. using biotic and a biotic factors has not been developed so far. To identify the biotic and a biotic factors that play a role in the grain yield and pest incidence the decision tree induction (DTI) and analysis in conjunction with Shannon information measure was explored. The developed classification model has the ability to successful treat 'categorical' variables as well as 'continuous' variables in the database. The information theoretic classification method used in the present study was aimed at finding a minimal set of database attributes involved in the induced model and was successful in grain yield or pest incidences. It was found that there was below 10% misclassified testing data. The confusion matrix for the testing set revealed that the classification was done more accurately using the training set. The developed prediction or classification model will helpful in forewarning about pest incidence and effect or grain yield and also to identify the factors influencing the pest population density. Using this model, agricultural farmers can apply pest control strategies on time to reduce crop loss.

Keywords: Sorghum Bicolor**,** decision tree, economic threshold level, classification, DTI, Multi location data, classification

IJERT

1. Introduction

Atherigona soccata Rondani (Muscidae: Diptera) commonly known as sorghum shoot fly is one of the most important pests in middle east, South east Asia, Africa and the Mediterranean Europe. It is serious pest infesting sorghum during *Kharif* and *Rabi* seasons in India. It is one of the major constraints in crop production. Under favourable conditions extent of damage up to 90 % has been reported. It is reported to cause economic loss to the tune of 120 million $. It causes maximum yield losses of 75.6% in grain and 68.6% in fodder. It causes round 22.4 – 39.5 % loss in grain yield. The damage is very high in late kharif sown crop as well as early rabi crop. On emergence, the neonate larvae crawl to the plant whorl and move downward between the folds of the young leaves. After reaching the growing point, it cuts the growing tip resulting in drying of the central leaf known as 'dead-heart' leading to the seedling mortality. The damage occurs 1 to 4 weeks after seedling emergence. The larval and pupal periods are completed in 8 to 10 days each.

The priority of research requires linking of pest incidence with weather parameters and to understand the contribution of natural enemies in pest incidence1. Data mining is a technique for extracting or mining of knowledge from a large database sets. Knowledge discovery (KD) is a nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in any data. The decision tree analysis in the data-mining technique is a popular predictive model classifier and predictor and is largely useful in classification applications, because it resembles human reasoning and can be easily understood. The decision can be represented in terms of a set of rules. The machine learning method relational decision tree model has been developed for the classification/prediction of pollen and seed dispersal of genetically modified (GM) and non-GM crops. The decision tree is a powerful tool for learning about the co-existence rules for GM and non-GM crops.

Knowledge-based systems were developed for solving problems concerned with pest identification, treatment prescription and strategic planning. Research has been carried on the development of user-friendly, computerized, expert systems for the management of pests and diseases in the Jamaican coffee industry and for crop protection in India and abroad considering factors like climate, topography, soil type of the farm, agronomic practices, crop phenology, biology and damage due to potential pests etc. However, there exists a gap in identifying the biotic and abiotic factors that play a role in the incidence of pest on different crops. According to Singh the main problem in addressing the issue of pest management is inadequate knowledge about the factors influencing the pest dynamics. The extent of damage may vary from crop to crop and season to season. Hence, the decision tree analysis for predicting pest incidence on sorghum crop, by analysing the abiotic and biotic factors has been proposed. The decision tree analysis in conjunction with Shannon information measure for the classification of pest incidence as high or low based upon economic threshold level (ETL) was explored for the occurrence of Shoot fly. The dataset was obtained from the D1Directorate of Sorghum Research (DSR), rajendranagar, Hyderabad -500 030 multi location experimental plots under All-India Coordinated

Sorghum Research Project (AICSIP) on Shoot flies. The sampling under weekly observations on mean number of shoot fly eggs damage present per plant was recorded for the period 20010-2013. Weather parameters like

IJERT

maximum temperature (Max T), minimum temperature (Min T), relative humidity (RH) and rainfall (RF) were taken based on weekly mean value. Pest incidence related with previous week biotic and a biotic factors was taken for decision tree analysis. The sample data are given in Table 1 and explanation of the attributes in the data base is given in Table 2. Pest incidence (PI) was the dependent variable (also referred to as class or target variable) and it was predicted based on several independent variables (also referred to as features or attributes). The overall dataset was divided into two parts, two thirds of the records were chosen for the training phase (from July 2010 to January 201) and one-third for the testing phase (during 2013).

A decision tree was built by initially selecting the instances from a training set. The intrinsic nature of the training set was then used by the algorithm to construct a decision tree for the testing set. Pest incidence was grouped into two classes, namely high and low based upon and pest incidence was considered as high if shoot fly dead hearts percentage more than 48 % and below the 48 % dead hearts low infestation recorded. The class values were assigned to the database based upon the user threshold input. The discretization technique, equal binning method was used to convert all numerical attributes (continuous variables) like PI, MaxT, MinT, RH and RF into categorical values/labels (A1–A5) as the option of total five bins was given. The labels of A1–A5 for each numerical attribute have their own range values stored in the database.

Table 1. Sample records/tuples from the database
Independent variable/feature/attribute Class/target

Tuple	P.I(dependent)	Max T °C	Min T°C	RH1(%)	RH2%	Rainfall (mm in the week)	Class
1	48.52	30.14	22.29	91.29	74.00	13.43	High
2	6.1	31.94	23.46	90.43	67.00	2.46	Low
3	13.9	28.37	22.11	99.14	76.86	8.26	Low
4	18.0	33.09	23.09	92.43	66.57	6.00	Low
5	18.0	22.9	14.1	35.4	17.0	0.5	Low

Table 2, Details of attributes: Attribute Explanation

PI	Pest incidence shoot fly dead hearts (%
MaxT	Maximum temperature (°C)
MinT	Minimum temperature (°C)
RH	Relative humidity (%)
RF	Rainfall (mm in the week)

IJERT

Materials Methods:

The mathematical representation of the binning method is as follows: (i) Get user input as the number of bins required (*b*); (ii) The number of bins *nb* = *n*/*b*, where *n* denotes the total number of records. (iii) Sort the numerical attribute values in ascending order. (iv) For each numerical attribute, assign the class label A1 into 1 to *nb* records, and A2 into *nb* + 1 to *nb* + *nb*, and A3 into 2*nb* to 2*nb* + *nb*, and so on, until the counter number of bins becomes zero. (v) The range values of A1 to A*n* was stored separately in the database for each numerical attribute. (vi) If extra records are available after creation of the bins, the last label, A*n*, was assigned.

Given a database *D* consisting of *t* data samples, where *D* = {*t*1, *t*2, …, *tn*} and a set of classes *C* = {*c*1, *c*2, …, *cm*}, the classification problem is to define a mapping *f* : *D* □ *C*, where each *ti* is assigned to one class. *D* contains *ti* tuples of class *Cj* for *j* = 1 to *m*, where *m* represents the number of classes and *m* = 2. The expected information needed to classify a tuple/record in *D* is given by:

$$\text{Info}(D) = \sum_{J=1}^{m} p_j \log_2(P_j)$$

where *pj* is the probability that an arbitrary tuple in *D* belongs to class *Cj* and is estimated using *n*(*Cj*, *D*)/*n*(*D*). Info(*D*) is also known as the entropy of *D*, i.e. total information value. Entropy of attribute (feature) *A* with values {*a*1, *a*2, …., *av*} is used to split *D* into *v* subsets:

$$\text{InfoA}(D) = \sum_{J=1}^{v} \frac{n(Dj)}{n(D)} \quad X \text{ Info}(Dj).$$

Info*A*(*D*) is the expected information required to classify a tuple from *D* based on the partitioning by *A*. Information is gained by branching an attribute *A*, and

Gain(*A*) = Info(*D*) – Info*A*(*D*).

Information gain is a measure of how good an attribute is for predicting the class of each of the training data. The attribute with the highest information gain was selected as the next split attribute according to the standard procedure[13.]

The attribute selection measure based on Shannon Information theory was used in the decision tree analysis. According to this, the maximum gain value attribute was chosen as the splitting attribute and based on the attribute the subsets were evaluated. Information measure was used recursively for each subset until the gain value or the entrophy reached zero for the attribute and this was used as a stopping criterion. The process continued until the search was completed and the attributes and labels stored in the table for tree generation. The model used the filter method as it selecting the features/attributes before applying an induction algorithm. The information theoretic method used in this model implemented automatic feature selection 'on the fly' as a part of the learning process. Thus, a minimal subset of features was found in a single run of the induction algorithm.

The root of the decision tree was fixed as PI and it was at level 0. The maximum gain value of the attribute was 'season' (Table 3). Hence, tree generation started from the season at level-1 and subsets had been generated automatically based on the information theory. The decision tree is a non-backtracking algorithm and hence it was constructed in a top-down manner.

The data-mining process for construction of the decision tree is given step by step in Figure 1. The training and testing process is given in Figure 2. The data obtained were subjected to regression analysis14 for developing the prediction equations for pest occurrence in sorghum crop based on biotic and biotic factors using the Statistical Package for Social Sciences15,

IJERT

ver. 17. The variable 'season' has been assigned numerical values 1–3 for monsoon, post-monsoon and winter seasons respectively.

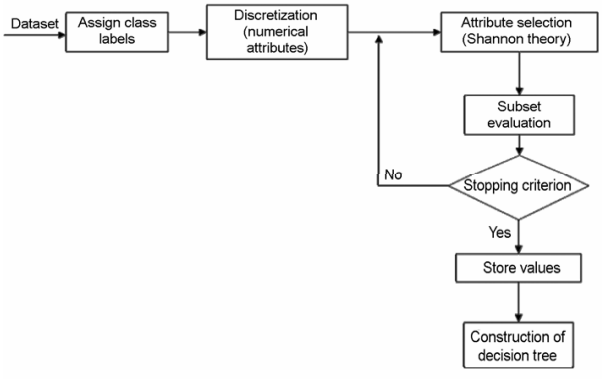


Figure 1. Data-mining process of construction of decision tree

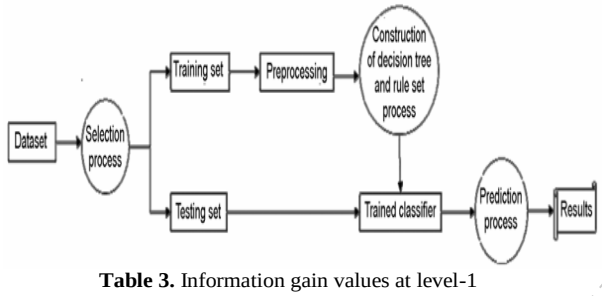


Table 3. Information gain values at level-1

Variable	Name	Gain value
PI	Pest incidence shoot fly dead hearts (%	0.27
MaxT	Maximum temperature (°C)	0.11
MinT	Minimum temperature (°C)	0.08
RH	Relative humidity (%)	0.03
RF	Rainfall (mm in the week)	0.01

The entrophy or information gain value at level-1 is given in the Table 3.

The constructed binary tree had decision node as the condition and output/result of that condition derived as the 'yes' or 'no' options. The 'yes' option always grows as left child and 'no' option as right child in the tree. The node ended when the condition was not able to proceed further in the 'no' option as right child. The end of the leaf node denoted the class label of pest incidence. The classes were always defined in the left child of the tree i.e. 'yes' option of the decision node. Root had been fixed as PI and the categorical variable season had been taken as the first attribute/variable starting from level-1 based on the highest gain value. The next attribute was elected based on the highest information gain value while in the process of feature subset selection as level-2, and so on until the information gain value became zero. The decision tree diagrams for *H. armigera* incidence during different seasons related with weather factors and natural enemies are given in Figures 3–5. IF–THEN rules derived from Figures 3 to 5 are given in Tables 4 for monsoon, post-monsoon and winter seasons respectively.

During monsoon season, when the maximum temperature ranged from 28.30°C to 30.80°C and minimum temperature ranged from 21.24°C to 22.59°C, pest incidence was high (Table 4). But, when maximum temperature ranged from 28.30°C to 30.80°C and NE1(spiders) was in the range 2.12–4.10, then pest incidence was low. Similarly, when the maximum temperature ranged from 30.81°C to 31.39°C, pest incidence was high when there was no rainfall (Table 4). But, when the maximum temperature ranged from 30.81°C to 31.39°C and rainfall ranged from 3.10 to 156.20 mm, the pest incidence was low (Table 5). Similarly, when the maximum temperature > 33°C with rainfall 0 mm, pest incidence was low, but when the maximum temperature > 33°C with rainfall in the range 18.40–156.20 mm, pest incidence was high (Table 5).

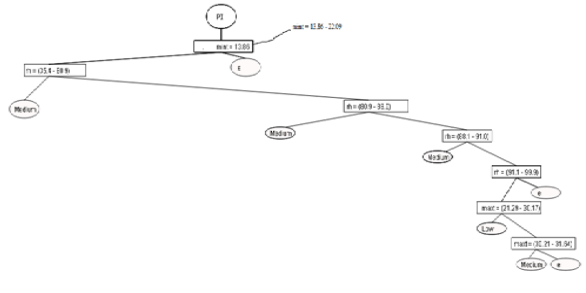
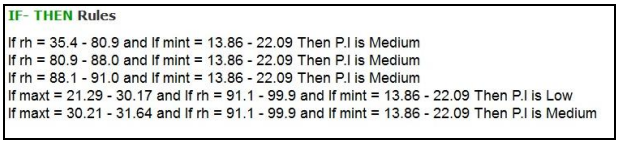


Table 4. Decision tree with numerical attributes categorized into five bins for Kharif.

IJERT


IF- THEN Rules

If rh = 35.4 - 80.9 and If mint = 13.86 - 22.09 Then P.I is Medium
If rh = 80.9 - 88.0 and If mint = 13.86 - 22.09 Then P.I is Medium
If rh = 88.1 - 91.0 and If mint = 13.86 - 22.09 Then P.I is Medium
If maxt = 21.29 - 30.17 and If rh = 91.1 - 99.9 and If mint = 13.86 - 22.09 Then P.I is Low
If maxt = 30.21 - 31.64 and If rh = 91.1 - 99.9 and If mint = 13.86 - 22.09 Then P.I is Medium

During post-monsoon season, when the maximum temperature ranged from 31.40°C to 32.14°C and from 28.30°C to 30.80°C, there was a greater chance of pest incidence (Table 5). Similarly, when minimum temperature ranged from 13.30°C to 15.87°C and from 15.94°C to 18.84°C, there was a greater chance of pest incidence. When minimum temperature ranged from 21.24°C to 22.59°C and from 22.59°C to 24.31°C, there was less chance of pest incidence during post-monsoon season Table 5). Similarly, when NE1 was in the range 2.12–4.10 per plant, pest incidence was low, which implied that spiders play a role in minimizing the pest occurrence (Table 5). During winter season, when the maximum temperature ranged from 28.30°C to 30.80°C and from 32.16°C to 33.31°C, the pest population was high. When the maximum temperature ranged from 33.40°C to 36.97°C, the pest population was low. But, when the maximum temperature
ranged from 33.40°C to 36.97°C and NE2 (*C. carnea*) ranged from 1.59 to 2.43 per plant, the pest population was low (Table 6). These findings of pest incidence during different seasons considering abiotic factors like maximum and minimum temperature were similar to those of earlier reports16–21. The above results show that when maximum temperature ranged from 33.40°C to 36.97°C, the pest population was low during monsoon and post-monsoon seasons. But, during monsoon with same range of maximum temperature and with rainfall, the pest population was high. During winter, with the maximum temperature in the same range, pest incidence was low as well as high. During monsoon and post-monsoon seasons, spiders played a role in reducing pest incidence, but during winter *C. carnea* played a role in reducing pest incidence. The prediction equation for *H. armigera* occurrence along with the coefficient of determination (*R*2) values and mean square error (MSE)22 are given in Table 6. The results of regression analysis revealed that about 66%, 21% and 40% of *H. armigera* incidence could be attributed to two natural enemies and the weather factors respectively (Table 6).
A comparison of Shannon information gain value, correlation analysis and regression analysis is given in Table 6

The Shannon information theory has shown that the attribute/variable or the factor 'Season' played an important role in pest incidence. The same had been proved with correlation analysis. Among the weather parameters, maximum temperature plays a major role in pest incidence because its *r* value is significant and also that its information gain value is more than other weather parameters values. Regression analysis also revealed that 'Season' played a major role in pest incidence among all other parameters (Table 5). The testing set was used for finding the accuracy of the classification. It was found that the misclassified testing
data were 8.82%. The confusion matrix derived for training set of data and testing set of data is given in Tables 5 and 6 respectively. The dataset for the period from July 2005 to January 2006 (training data) used for training the model such that assignment of classes based on PI threshold values, categorization of numerical attributes based on the binning method, storage of range values for the numerical attributes, finding information gain table and construction of decision tree. The dataset for the period from August 2007 to January 2008 (testing data) was given to the model for testing. In the testing phase, the class values assigned for PI, categorization of numerical attributes carried out based on training data information. Then, the information gain table was calculated for testing records and

IJERT

the decision tree was constructed. Confusion matrix was derived for the testing the records after the classification process and the percentage of misclassification was 8.82. Hence, the testing set revealed that prediction or classification of pest incidence was done 91.18% accurately. Moreover, the decision tree visualized the biotic and abiotic factors with range values playing a role in pest incidence (either 'LOW' or 'HIGH'). The occurrence of shoot fly was greatly influenced by its natural enemies, viz. spiders and *C. carnea* and by abiotic factors. In the present work, the population dynamics of the pest and its natural enemies was studied using Shannon information measure with decision tree induction approach. The developed classification model has the ability to successfully treat 'categorical' variables as well as 'continuous' variables in the database. Pest incidence had been classified

Table 5. Prediction equation for Shoot fly dead hearts %

Y = 5590 + 1.268 * Max Temp oC – 3.691 * Min Temp oC – 0.119 * Rh1 (%) + 0.354 * Rh2 (%) – 0.107 * Rain fall (mm) for Low
$R^2 = 0.4452$
MSE = 0.462

Max Temp = Maximum Temperature; Min temp = minimum Temperature; Rh1 = Relative humidity1; Rh2 = Relative humidity2and Rain fall in mm week days.

Table 6. Comparison of information gain values, correlation analysis and regression analysis

Attribute	Information gain value	Correlation analysis (r)	Regression analysis
Min Temp	0.27	-0.44 **	Y = 5590 + 1.268 * Max Temp oC – 3.691 * Min Temp oC – 0.119 * Rh1 (%) + 0.354 * Rh2 (%) – 0.107 * Rain fall (mm)
Max Temp	0.11	-0.41**	$R^2 = 0.4452$
Rh1 (%)	0.08	-0.01	
Rh2 (%)	0.03	-0.03	
Rain fall	0.01	0.1	

Conclusion:

This Decision analysis approach could be successfully understand the role of pest incidence and prediction. This analysis is suitable classifier model to predict pest incidence in more than 90% accuracy since tree analysis is simple and easy to understand and It can apply many other discipline traits of sorghum crop since it don't require domain knowledge

References

1. DSR, Annual Reports of AICSIP group meeting (AGM,12), Rajendra nagar, Hyderabad
2. Raghuram Raghava Sharma, Data ware housing and data mining, June, 20o11, First edition, Professional publications, JNTU, Hyderabad.
3. Introducttion to data mining , Sixth impression, 2011, Pang Ning Tan, Vipin Kumar and Michael Stein bach.
4. Data mining concepts and techniques, 2012, Elseveir, Thord Edition, Jiawei, Michiline Kamber and Jian Pei.
5. Jaiwei Han et al. 2012, Elsevier, Data mining concepts and techniques.
6. Ravindra Changala et al. Volume 2, Issue 4, April 2012 ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering, Classification by Decision Tree Induction Algorithm to Learn Decision Trees from the class-Labeled Training Tuples,
7. Kelley and Parthasarathy Rao 1994, Hall 2000, "An analysis of availability and utilization of sorghum grain in India, ICRISAT, India, of availability and utilization of sorghum grain in India". Journal of SAT Agricultural Research.
8. Rokach, Lior; Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc. ISBN 978-9812771711.
9. Deng,Hm et al. 2011 . "Bias of importance measures for multi-valued attributes and solutions". Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN). pp. 293–300.

IJERT

10. Trivedi, T. P., Yadav, C. P., Vishwadhar, Srivastava, C. P., Dhandapani, A., Das, D. K. and Singh, Joginder, Monitoring and forecasting of *Heliothis*/*Helicoverpa* population. In *Heliothis/ Helicoverpa Management – Emerging Trends and Strategies for Future Research*, 2005, pp. 119–140.
11. Zhao, H. and Ram, S., Constrained cascade generalization of decision trees. *IEEE Trans. Knowledge Data Eng*., 2004, **16**(6), 727–739.
12. Basak, J. and Krishnapuram, R., Interpretable hierarchical clustering by constructing an unsupervised decision tree. *IEEE Trans. Knowledge Data Eng.*, 2005, **17**(1), 121–132.7.
13. Aneta, I., Celine, V., Nathalie, C., Marko, D. and Saso, D., The feasibility of co-existence between conventional and genetically modified crops: using machine learning to analyse the output of simulation models. *Ecol. Model*., 2008, **215**, 262–271.
14. Chakraborty, P. and Chakrabarti, D. K., A brief survey of computerized expert systems for crop protection being used in India. *Prog. Nat. Sci.*, 2008, **18**(4), 469–473.
15. Mansingh, G., Reichgelt, H. and Osei Bryson, K. M., CPEST: An expert system for the management of pests and diseases in the Jamaican coffee industry. *Expert Syst. Appl*., 2007, **32**(1), 184– 192
16. Gupta, G. K., Classification. In *Introduction to Data Mining with Case Studies*, Prentice-Hall of India, 2006, pp. 106–136.
17. SPSS V 17.0, Statistical Package for Social Sciences, SPSS Inc.,
18. Illinois, Chicago, USA, 2008
19. Dhaliwal, L. K., Kooner, B. S., Singh, J. and Sohi, A. S., Incidence of *Helicoverpa armigera* (Hübner) in relation to meteorological parameters under Punjab conditions. *J. Agrometeorol*. (*Spec. Issue*), 2004, **6**, 115–119.

IJERT