A Decision Tree Induction Approach to Study the Population Dynamics of *Helicoverpa armigera* (Hübner) and its Natural Enemies on Cotton

Meena, K[1]., Pratheepa, M[2*]., Subramaniam, K. R[3]., Venugopalan, R[4] and Bheemanna H[5].

Abstract - **In this present communication, effort has been made to study the population dynamics of *H. armigera* (Hübner), one of the agriculturally important crop pest and its natural enemies' *viz*., spiders and chrysopids on cotton crop by using decision tree induction approach. It has been found that the percentage of misclassified training data is 20% and the total time taken to generate the decision tree is 22 sec. The confusion matrix for testing set revealed that the classification was done more accurately by using the training set. Hence, it is proved that this approach can be successfully utilized to understand the role of natural enemies and weather factors on the population dynamics of *H. armigera* (Hübner). Our studies showed that maximum temperature influenced the population dynamics of *H. armigera*. This classification model will help the farmers to make preparations in time and to decide the timing application of pest control strategy.**

Index Terms - **Data mining, Decision tree classification, Pest prediction, Statistical models**

I. INTRODUCTION

The cotton bolloworm *(Helicoverpa armigera*) is one of the most important constraints to crop production globally [9]. Over the past two decades, *H.armigera* damage has resulted in tremendous crop loss, particularly in cotton and pulse crops [8]. Frequent outbreaks of *H. armigera* are common in India leading to various social and economical problems. Using long-term climatic and insect population data, it is now possible to predict the chances for pest build up in the crop. Although the current models do not predict the pest occurrences properly, there is a good scope to use the emerging 'Information and Communication Technology (ICT)' to accurately predict pest build up. This can then be used as an easily warning system to advice farmers for appropriate management practices to minimize damage by *Helicoverpa.* Some of the techniques like statistical, mathematical and simulation models have been developed on forecasting the pest *Helicoverpa armigera* (Hübner) [8]. But, still there is a gap on findings of the role of weather factors and natural enemies on pest occurrence. Single model is insufficient to forecast *H. armigera* population all over the entire continent [8]. Hence, the efforts had been made to study the population dynamics of *H.armigera* on cotton crop by using decision tree induction approach.

II. DECISION TREE

A decision tree is a popular classification method and is largely popular in classification applications because models they generate closely resemble human reasoning and are easily understood [2]. The major advantage of the decision tree is its interpretability, i.e. the decision can be represented in terms of set of rules [3]. The decision tree algorithm is a relatively simple non-backtracking and constructed in a top-down manner [1].

Classification is a type of prediction problem where the dependent variable (also referred to as class) that needs to be predicted with greater accuracy based on several independent variables (also referred to as features and attributes) are categorical and continuous [2]. A learning technique constructs a hypothetical model, which is a mapping from the independent variables to the dependent variable, by investigating a given set of successfully solved cases, whose output on the dependent variable are already known; the model can then be used to predict the outputs of unseen cases on the dependent variable. In the proposed algorithm, the pest incidence (PI) or pest density is taken as dependent variable and the other attributes like season, number of Spiders per plant (NE1), number of Chrysopids per plant (NE2), Maximum temperature (^{0}C), Minimum temperature (^{0}C), amount of Rain fall and Relative humidity are acting as independent variables. A n-class classification problem is described by a pair <X,Y>, where $X = X_1 \text{ x } X_2 \text{ x } \ldots \text{ x } X_m$ are an m-dimensional space (independent variables) and Y = { 1,2, …, n} is a discrete space, where n varies from 1 to 4. The pest incidence (Y=PI) is classified into 'n' number of classes like high, medium, low and when there is no pest incidence it is described as nil class.

III. DATA BASE

The data set were obtained from Regional Agricultural Research Station, Raichur, Karnataka, India from the unsprayed experimental plots under AICRP on NCS-145 Non Bt-Cotton. Weekly observations on mean

* This is part of Ph.D work by the second author

[1] Principal & Director of Shrimathi Indira Gandhi College(SIGC),Trichirapalli,Tamil Nadu,INDIA.

[2] ☞Corresponding author,Ph.D Scholar, SIGC,Trichirapalli,Tamil Nadu, INDIA.

[3] Professor & Head, Dept. of M.C.A.,SIGC,Trichirapalli, Tamil Nadu, INDIA.

[4] Senior Scientist(Agril.Statistics),Indian Institute of Horticultural Research,ICAR,Bangalore,INDIA.

[5] Senior Scientist(Entomology),UAS,Agricultural Research Station,Karnataka,INDIA.

number of *H. armigera* larvae present per 5 plants were recorded for the period 2005 to 2008. Larval counts then were converted to per plant for the pest incidence (PI) and natural enemies' (NE) population *viz.*, spiders and chrysopids recorded per plant. The weather parameters like maximum temperature (MaxT), mimimum temperature (MinT), Relative humidity (RH) and Rainfall (RF) were taken based on weekly mean value. The sample data set are given in Table I and the explanation for each of the attribute/variable is given in Table II.

Table I. Sample data set of pest incidence (PI), natural enemies (NE), temperature, rainfall (RF) and relative humidity (RH) at two different seasons (Monsoon and Post Monsoon)

PI	Season	NE1	NE2	MaxT	MinT	RF	RH
0.65	Monsoon	0.39	0.20	32.2	21.2	46.8	71
0.45	Monsoon	0.41	0.26	30.2	22.6	81.4	77
0.7	Post Monsoon	0.45	0.22	31.1	21.5	1.4	68
0.58	Post Monsoon	0.46	0.28	28.3	17.8	21.4	76
0.65	Post Monsoon	0.47	0.31	30.1	16.3	0.0	59
0.8	Post Monsoon	0.52	0.36	30.2	13.4	0.0	57
1.06	Post Monsoon	0.58	0.39	30.5	16.3	0.0	66
1.15	Post Monsoon	0.59	0.49	29.9	16.5	0.0	60
0.62	Post Monsoon	0.62	0.58	31.0	17.1	0.0	55

Table II. Details of attributes

Attribute	Explanation
PI	Pest incidence per plant
Year	Observation taken on that year
Season	Observation taken on this season
NE1	Natural Enemy 1 (No. of Spiders per plant)
NE2	Natural Enemy 2 (No. of Chrysopids per plant)
MaxT	Maximum Temperature in Celsius
MinT	Minimum Temperature in Celsius
RH	Relative humidity in percentage
RF	Rainfall in mm

IV. DATA PREPROCESSING

Data preprocessing involved data selection, data reduction and removal of null values. All these were carried out on the data set and missing values were filled with mean values.

V. TRAINING PHASE AND TESTING PHASE

The overall data set were divided into two parts. *ie.*, 2/3rd of records had chosen for training phase and 1/3rd of records had chosen for testing phase.

VI. ASSIGNING CLASS LABEL

The pest incidence was classified into four classes namely nil, low, medium and high. The arbitrary user given threshold values had been used in the training phase to assign the class label for each record. The records had been trained in the training phase for assigning class labels like low, medium and high. Nil class assigned where the pest incidence not occurred. The class label attribute was discrete-valued and unordered. The maximum and minimum values used to get the range values for continuous independent variables.

VII. INFORMATION THEORY

The data classification process is aimed at reducing the amount of uncertainty or gaining information about the target (classification) attribute. The Shannon theory, information gain is used as attribute selection measure. It is a measure of how good an attribute is for predicting the class of each of the training data [1]. The concept is quite simple. It is based on Claude Shannon's idea that if you have uncertainty then you have information and if there is no uncertainty there is no information [1]. The expected information needed to classify a tuple/record in D is given by

$$\text{Info (D)} = \sum_{i=1}^{m} p_i \log_2(p_i)$$

where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $n(C_i,D)$ / n(D). Info(D) is also known as the entropy of D. *ie.*, total information value [4]. Let D, the data partition, be a training set of class labeled tuples. Suppose the class label attribute has 'm' distinct values defining 'm' distinct classes, C_i (for I = 1, 2,…,m). Let C_i,D be the set of tuples of class C_i in D. Let $n(C_i,D)$ denotes total no. of tuples or records in class C_i and n(D) denotes total no. of tuples or records in D respectively.

$$\text{Info}_A(\text{D})= \sum_{j=1}^{v} \frac{n(D_j)}{n(D)} xInfo(D_j)$$

Info_A is denoted as the information value for the attribute A.. Info_A(D) is the expected information required to classify a tuple from D based on the partitioning by A.

Gain value for an attribute A is calculated as

Gain (A) = Info (D) – Info_A (D)

Gain (A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, (Gain (A)), is chosen as the splitting attribute at node N. Here, the attribute A would do the “best classification” [4]. Information gain table given a goodness value that shows the discrimination power of the attribute. The attribute then may be ranked according to that value and the highest ranked attribute selected as level 1 and the remaining attributes are selected as level 2, etc. The highest value of information gain value of an attribute gives information that this variable plays major role on predictor variable.

PI variable or an attribute was divided into 3 classes based upon the user defined threshold ranges as pest incidence is low, medium and high. NIL also defined when there was no occurrence of pest incidence. As on total, the number of classes ranged from i = 1 to 4, where m = 4.

VIII.INFORMATION GAIN VALUE

Information gain value was calculated for other independent variables like NE1, NE2, MaxT, MinT, RF and RH. It had been found that MaxT had the highest gain value and inferred that Maximum temperature played major role on pest incidence than other weather parameters. The gain value for NE1, NE2 derived as null value, which denotes that natural enemies were not playing major role on pest incidence/density. The gain values are depicted and ranked in Table III.

Table III. Information gain table for the independent variables

Attribute	**Gain value**	**Rank**
MAXT	0.87956669	1
RF	0.73259095	2
MINT	0.18693437	3
RH	0.16256725	4
NE2	0	
NE1	0	

IX. CONSTRUCTION OF DECISION TREE

Based on user threshold values the class labels assigned and the maximum-minimum values used to find out the range for continuous data attributes. Mode value had been chosen for categorical attributes. The tree constructed in the form of binary tree. The tree has decision node as condition and output/result of that condition derived as yes or no options. The ‘yes’ option always grows as left child and ‘no’ option always grows as right child in the tree. The node ends when the condition not able to proceed further in the ‘no’ option as right child. The end of the leaf node denotes the class label of pest incidence. The classes always defined in the left child of the tree ie. ‘yes’ option of the condition/decision node. Root had been fixed as PI and the categorical variable season had been taken as the first attribute/variable starting from the level 1. The next attribute had been selected based on the ranking of information gain table as level 2, and so on. The decision tree diagram for *H. armigera*(Hübner) related with weather factors is given in Fig.1 and decision tree diagram for *H. armigera*(Hübner) related with natural enemies is given in Fig. 2.

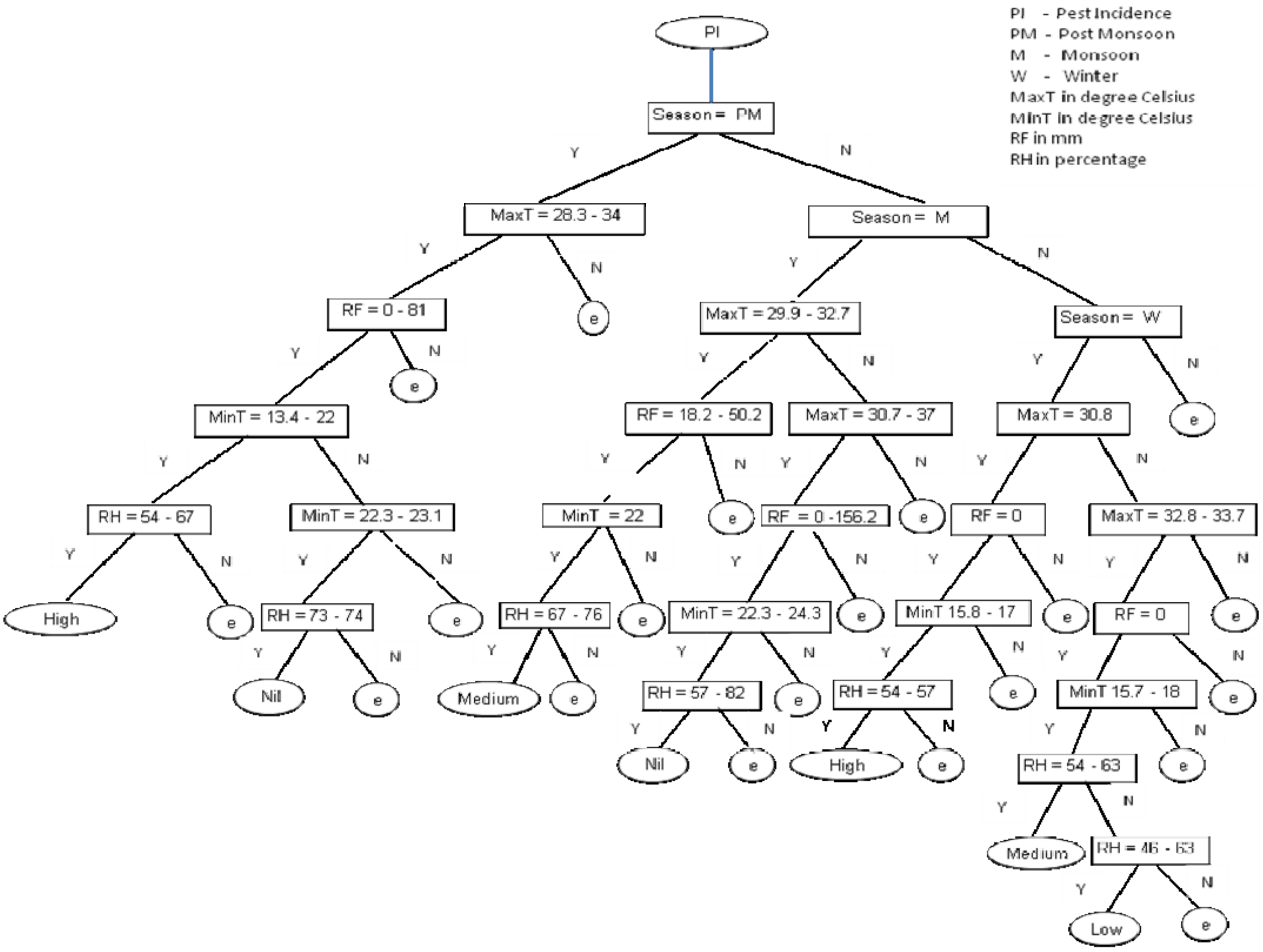

Fig 1. Decision tree for training set of data (PI with weather parameters)

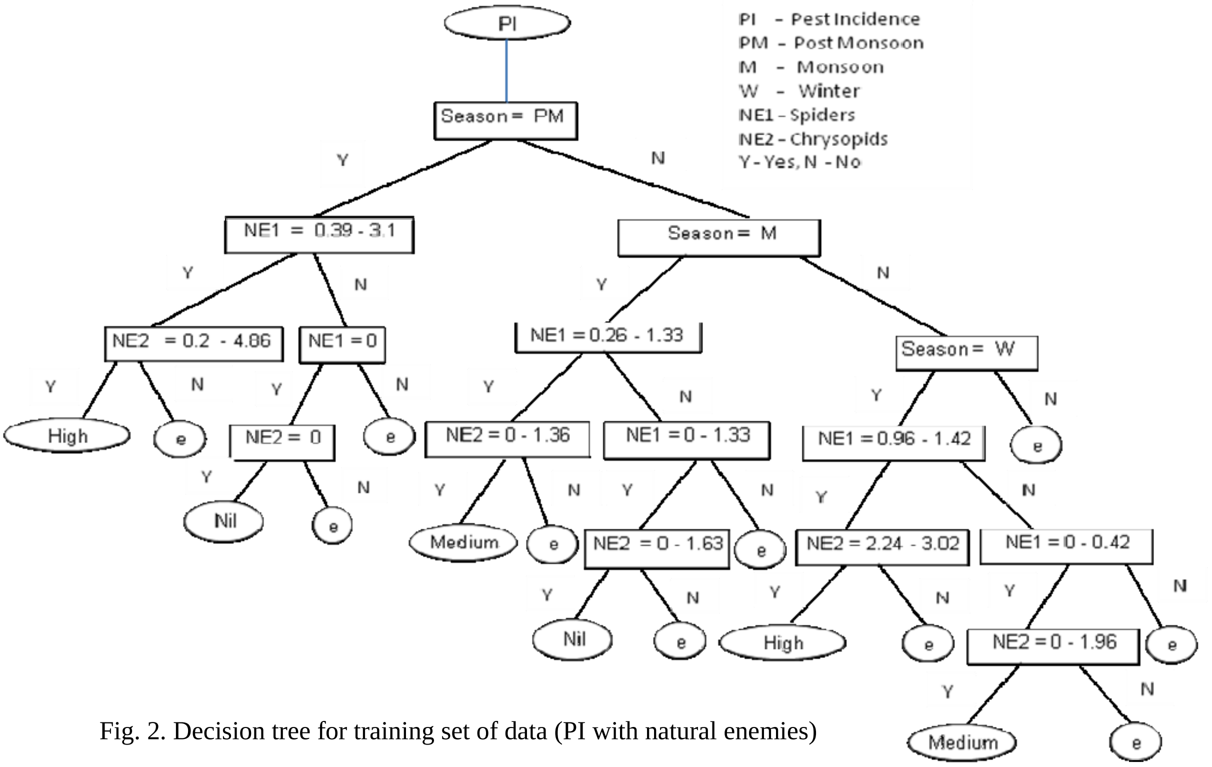

Fig. 2. Decision tree for training set of data (PI with natural enemies)

X. KNOWLEDGE DISCOVERY

IF – THEN rules derived by using the decision tree in Fig. 1 and the rules for PI related with weather factors are given in Table IV. IF – THEN rules derived by using the decision tree in Fig. 2 and the rules for PI related with natural enemies are given in Table V.

Table IV. Set of rules for pest incidence (PI) related with weather factors

Condition or rule or decision			**Result /Class (PI)**
IF	Season=Post monsoon and MaxT = 28.3°C - 34°C and RF = 0 – 81 mm and MinT 13.4°C – 22.6°C and RH = 54 - 77%	THEN	High
IF	Season=Post monsoon and MaxT = 28.3°C - 34°C and RF = 0 – 81 mm and MinT = 22.3 – 23.1 °C and RH = 73 - 74%	THEN	Nil
IF	Season=Monsoon and MaxT = 29.9°C – 32.7°C and RF = 18.2 – 50.2 mm and MinT = 22°C and RH = 67 - 76%	THEN	Medium
IF	Season=Monsoon and MaxT = 30.7°C – 37°C and RF = 0 – 156.2 mm and MinT = 22.3 – 24.3°C and RH = 57 - 82%	THEN	Nil
IF	Season=Winter and MaxT = 30.8°C and RF = 0 mm and MinT = 15.8 - 17°C and RH = 54 - 57%	THEN	High
IF	Season=Winter and MaxT = 32.8 – 33.7°C and RF = 0 mm and MinT = 15.7 - 18°C and RH = 54 - 63%	THEN	Medium
IF	Season=Winter and MaxT = 32.8 – 33.7°C and RF = 0 mm and MinT = 15.7 - 18°C and RH = 46 - 63%	THEN	Low

The IF-THEN rules from Table IV reveals that when maximum temperature > 34˚C and minimum temperature > 22˚C, then there is less chance of pest occurrence. Similarly when MaxT ranges from 28˚C - 34˚C and MinT 13˚ - 22˚C then there is greater chance of pest occurrence.

The IF-THEN rules from Table V reveals that there is no much role of natural enemies spiders and chrysopids on subsiding of the pest incidence *H. armigera*(Hübner).

Table V. Set of rules for pest incidence (PI) related with natural enemies

Condition or rule or decision			**Result /Class (PI)**
IF	Season = Post monsoon and NE1 = 0.39 – 3.1 and NE2 = 0.2 – 4.86	THEN	High
IF	Season = Post monsoon and NE1 = 0 and NE2 = 0	THEN	Nil
IF	Season = Monsoon and NE1 = 0.26 – 1.33 and NE2 = 0 – 1.36	THEN	Medium
IF	Season = Monsoon and NE1 = 0 – 1.33 and NE2 = 0 – 1.63	THEN	Nil
IF	Season = Winter and NE1 = 0.96 – 1.42 and NE2 = 2.24 – 3.02	THEN	High
IF	Season = Winter and NE1 = 0 – 0.42 and NE2 = 0 – 1.96	THEN	Medium

XI. RESULTS AND DISCUSSION

To understand the role of natural enemies and weather factors on pest incidence, a statistical model relating PI vs all factors were developed and the results are presented in Table VI.

Table VI. Correlation coefficients of pest incidence with natural enemies and weather factors based on class-wise classification

	High	**Medium**	**Low**	**Over all**
	PI	**PI**	**PI**	**PI**
NE1	0.05	-0.27	-0.35	0.08
NE2	**0.32**	-0.10	-0.37	**0.36**
MaxT	0.04	0.05	-0.49	**-0.32**
MinT	-0.07	0.11	-0.05	**-0.27**
RF	-0.30	-0.13	0.27	**-0.26**
RH	0.05	0.23	-0.12	**-0.24**

r value bolded are significant at $P<0.05$

When population density of *H.armigera* was high, chrysopids (NE2) exhibited significant inference on pest population. While at medium and low pest density none of biotic and abiotic factor played any significant role on pest population. Irrespective of pest population, chrysopids exhibited positive correlations with pest density, whereas abiotic factors *viz.* maximum temperature, minimum temperature, rain fall and relative humidity led significant negative correlation with pest density.

The data was also subjected to regression analysis [6] and the equations along with their Co-efficient of determination (R^2) values and Mean Square Error (MSE)

[5] are given in Table VII. Statistical Package for Social Science (SPSS V 17.0) was extensively used in model building exercise.

Table VII. Results of Regression analysis

Case/dependent variable	Full Model	R^2	MSE
High class	Y= -0.46 - 0.08 NE1+0.06 NE2 +0.02MaxT – 0.03 MinT – 0.005RF + 0.02 RH	0.35	0.04
Medium class	Y= -0.6 - 0.07 NE1+0.03 NE2 +0.002MaxT – 0.006RF + 0.02 RH	0.73	0.002
Low class	Y= 0.51 - 0.01 NE1- 0.01 NE2 - 0.01MaxT – 0.005 MinT + 0.0001RH	0.38	0.001

Over all
A) Full regression model
B) Optimized model
	Y= 3.9 - 0.11 NE1+0.17 NE2 - 0.09MaxT + 0.0001RF - 0.009 RH
	Y= 4.24 + 0.11 NE2 - 0.1MaxT - 0.01 RH
		0.48
		0.43
			0.064
			0.07

Where Y= *Helicoverpa armigera* , NE1=Spiders, NE2=Chrysopids, MaxT = Maximum temperature, MinT = Minimum temperature, RF = Rainfall, RH = Relative humidity.

The results of regression analysis revealed that about 35%, 73%, 38% and 48% of the population dynamics of *H.armigera* could be attributed to two natural enemies and the weather factors, respectively.

The variables/attributes MaxT, MinT, RF are negatively correlated with pest incidence/density and season is positively correlated with pest incidence. The information gain table also reveals that among all weather parameters maximum temperature gain value is the maximum and it exhibits that there is strong relation between the pest incidence and the maximum temperature. The multiple regression optimized model also reveals that MaxT is negatively correlated with pest incidence which means of the 1% decrease of MaxT will give chance of increasing the pest incidence.

The comparison of Shannon information gain value, correlation analysis and regression analysis are given in Table VIII.

Table VIII. Comparison of Information gain values, Correlation analysis and Regression analysis

Attribute name	Information Gain Value	Correlation analysis	Regression analysis
MaxT	**0.88** (maximum)	**-0.32**	Y= 3.9 - 0.11 NE1+0.17 NE2 - 0.09MaxT + 0.0001RF - 0.009 RH **(R^2 = 48% and MSE = 0.064)**
RF	0.73	**-0.26**	
MinT	0.19	**-0.27**	
RH	0.16	**0.36**	
NE2	0	0.08	
NE1	0		

r value bolded are significant at $P < 0.05$

The Shannon Information Theory had been proved that the attribute/variable or the factor MaxT was playing a most important role for the cause of pest incidence. The same had been proved with correlation analysis. Regression analysis also revealed that maximum temperature played a major role on pest incidence among all other weather parameters.

The confusion matrix had been derived for the training set and testing set and it has been depicted in Table IX and Table X.

Table IX. Confusion matrix for Training Data (2/3 of total records)

True Class	Predicted class				
	High	Medium	Low	Nil	Total
High	24				24
Medium		11			11
Low		1	3		4
Nil		2		10	12
Total	24	14	3	10	51

Table X. Confusion matrix for Testing Data (1/3 of total records)

True Class	Predicted class				
	High	Medium	Low	Nil	Total
High	2				2
Medium	2				2
Low	9	1	1		11
Nil		5	1	2	8
Total	13	6	2	2	23

XII. CONCLUSION

Pest population density can be strongly influenced by biotic and abiotic factors. These factors interact and are usually auto correlated and hence it is difficult to separate one variable from the other. In this present work, the population dynamics of *H. armigera* (Hübner) and its natural enemies' *viz.*, spiders and chrysopids on cotton has been studied thoroughly by using decision tree induction approach. It had been found that the percentage of misclassified training data was 20% and the total time taken to generate the decision tree was 22 sec. The confusion matrix for testing set revealed that the classification was done more accurately by using the training set. Hence, it is proved that this approach can be successfully utilized to understand the role of natural enemies and weather factors on the population dynamics of *H. armigera* (Hübner). Our studies showed that maximum temperature influenced the population dynamics of *H. armigera*. This classification model will help the farmers to make preparations in time and to decide the timing application of pest control strategy.

REFERENCES

[1] G. K. Gupta, "Classification", in *Introduction to Data Mining with Case Studies*, Prentice-Hall of India, 2006, pp.106–136.

[2] H. Zhao and S. Ram, "Constrained cascade generalization of decision trees", *IEEE Transactions on Knowledge and Data Engineering.* Vol.16(6), 2004, pp.727-739.

[3] J. Basak and R. Krishnapuram, "Interpretable hierarchical clustering by constructing an unsupervised decision tree", *IEEE Transactions on Knowledge and Data Engineering*. Vol.17 (1), 2005, pp. 121-132.

[4] Jiawei Han and Micheline Kamber, "Classification and Prediction" in *Data Mining Concepts and Techniques*, 2nd ed. Jim Gray Ed. 2001, pp. 285-375.

[5] T. O. Kvalseth, "Cautionary note about R2", *The American Statistician.* Vol. 39, 1985, pp. 279-285.

[6] Ryan, P. Thomas, *Modern Regression Methods*. John Wiley and Sons, Inc, New York, 1997.

[7] SPSS V 17.0, Statistical Package for Social Sciences. SPSS Inc. Chicago, Illinois, USA, 2008.

[8] T. P. Trivedi, C. P. Yadav, Vishwadhar, C. P. Srivastava, A. Dhandapani, , D. K. Das and Joginder Singh, "Monitoring and forecasting of *Heliothis/Helicoverpa* population", in *Heliothis/Helicoverpa Management – Emerging Trends and Strategies for Future Rresearch*. 2005, pp. 119-140.

[9] M. P. Zalucki, G. Daglish, S. Firempong and P. Twine, "The biology and ecology of *Heliothis armigera* (Hübner) and *H.punctigera* Wallengren (Lepidoptera:Noctuidae) in Australia: What do we know?", *Australian Journal of Zoology*, Vol. 34, 1986, pp. 779-814.