Journal of Biological Control, 32(1): 31-36, 2018, DOI: 10.18311/jbc/2018/18163

Research Article

Optimized binning technique in decision tree model for predicting the *Helicoverpa armigera* (Hübner) incidence on cotton

M. PRATHEEPA[1*], J. CRUZ ANTONY[2], CHANDISH R. BALLAL[1] and H. BHEEMANNA[3]
[1]*ICAR-National Bureau of Agricultural Insect Resources, Bengaluru – 560024, Karnataka, India*
[2]*Department of Computer Science, Jain University, Bengaluru – 560011, Karnataka, India*
[3]*University of Agricultural Sciences, Agricultural Research Station, Raichur - 584102, Karnataka, India*
[*]*Corresponding author E-mail:* mpratheepa@gmail.com

ABSTRACT: The data mining technique decision tree induction model is a popular method used for prediction and classification problems. The most suitable model in pest forewarning systems is decision tree analysis since pest surveillance data contains biotic, abiotic and environmental variables and IF-THEN rules can be easily framed. The abiotic factors like maximum and minimum temperature, rainfall, relative humidity, etc. are continuous numerical data and are important in climate-change studies. The decision tree model is implemented after pre-processing the data which are suitable for analysis. Data discretization is a pre-processing technique which is used to transform the continuous numerical data into categorical data resulting in interval as nominal values. The most commonly used binning methods are equal-width partitioning and equal-depth partitioning. The total number of bins created for the variable is important because either large number of bins or small number of bins affects the accuracy in results of IF-THEN rules. Hence, optimized binning technique based on Mean Integrated Squared Error (MISE) method is proposed for forming accurate IF-THEN rules in predicting the pest *Helicoverpa armigera* incidence on cotton crop based on decision tree analysis.

KEY WORDS: Bin optimization, decision tree, discretization, *Helicoverpa armigera*, IF-THEN rules, pest prediction

(Article chronicle: Received: 14-11-2017; Revised: 26-02-2018; Accepted: 10-03-2018)

INTRODUCTION

There are several data mining techniques like logistic regression, decision tree analysis, Bayesian Networks (BNs) and Rule-Learners (RLs) which are widely used in prediction models. These techniques otherwise called as classification techniques. Classification techniques used to predict the target variable in the form of categorical class labels as "Yes/No", "Present/Absent", "High/Low", etc. Classification algorithm uses nominal values of independent variables to predict their class labels of target variable. Generally independent variables are in the form of real-valued attributes or continuous numerical values as in the case of weather data which contains maximum temperature, minimum temperature, relative humidity, rain fall, etc. Suitable discretization algorithms are needed to handle problems of conversion of real valued attributes of independent variables to nominal or categorical values.

Decision tree analysis and rule-learners could be suitable model for pest prediction by using weather data, since the IF-THEN rules derived from these models are easily understandable and interpretable (Zhao and Ram, 2004). Discretizing real-valued continuous attributes is an important technique in data pre-processing to select the relevant features in classification algorithms. Discretization is usually performed prior to the induction or learning process. George *et al.,* (1994) addressed about the relevant features and irrelevant features selection in decision tree algorithm ID3. In discretization process the continuous element array has been divided into different bins/buckets/intervals. The term "cut-point" refers to the point where the partition occurs in an array. Let us consider a continuous interval [a,b] is partitioned into [a,c] and [c,b] where 'c' is known as cut-point or split-point (Sotiris and Dimitris, 2006). The cut-point or split-point has been chosen based on the differ-

ent methods and call it as binning methods. There are several unsupervised binning methods available and they are (i) Equal-width binning (ii) Equal-frequency binning (iii) Max-diff methods. Decision tree analysis with equal binning discretization method resulted ambiguity in IF-THEN rules (Pratheepa *et a*l., 2011). The total number of decision tree levels depends upon the number of intervals or bins, partitioned in an attribute. A good discretization algorithm is needed to generate reasonable number of cut points (Sotiris and Dimitris, 2006). Hence the optimized binning method has been proposed to avoid ambiquity and to derive optimized IF-THEN rules in *Helicoverpa armigera* prediction on cotton crop based on biotic and abiotic factors. Comparison has been done for the decision tree analysis with equal binning method and with optimized binning method and found that the optimized binning method gives better performance than the conventional equal binning method.

MATERIALS AND METHODS

Data collection

The data set were obtained from the Regional Agricultural Research Station (16°.21'N/77°.34'E), Raichur, Karnataka from the unsprayed experimental plots under All-India Coordinated Research Project (AICRP) on non-Bt-cotton. The sampling size was 25 plants/500 m^2 areas. Weekly observations were made on the mean number of *Helicoverpa armigera* larvae present per 25 plants for the period 2005-2013. The pest incidence with mean values of previous one week weather parameters *viz*., MaxT, MinT, RH1, RH2 and total RF and RFD (No. of rainfall days) of previous one week, biotic factors *viz*., spiders (NE1) and coccinellids (NE2) per plant were taken for analysis. The sample data set has been given in Table 1. The crop stage of cotton crop and season as per the India Meteorological Department (IMD) of southern India have been considered along with this data set. There were 5 crop stages on cotton crop like (i) 1–5 weeks of the plant (ii) square initiation stage (iii) flowering and boll formation stage (iv) boll maturity stage and (v) boll bursting stage and numbered as 1 to 5.

Assign class label to the target variable

To carry out decision tree analysis the target variable 'PI' is to be converted into the class values like 'High' or 'Low' based on Economic Threshold Level (ETL) of *H. armigera* on cotton. When the pest incidence was above 1 larva/10 plants, it was considered as 'High' and when the pest incidence was below 1 larva/10 plants, it was considered as 'Low' as mentioned in Table 1 (Dhaliwal and Arora, 1996).

Discretization of continuous variables

It is necessary to convert the continuous values of weather parameters MaxT, MinT, RH1, RH2, RF and biotic factors NE1 and NE2 into categorical or nominal values to implement decision tree analysis to derive IF-THEN rules. The bin optimization method has been adopted based on (Shimazaki and Shinomoto, 2007) in this paper.

The steps involved in bin optimization method are given below:-

Step1: Sort the array in ascending order // eg. MaxT
Step 2: icost, c(Δ) = 0 // Initialize cost value
Step 3: For nb = 2 to 6 // No. of bins initialized as 2 to 6
Step 4: The mean and variance of each bin is calculated with the following formulae (i) and (ii) respectively.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(i)}$$

Table 1. Sample data set for *Helicoverpa armigera* incidence with biotic and abiotic factors

PI	Class	Year	Cropstage	Season	Maxt (°C)	Mint (°C)	RH1 (%)	RH2 (%)	RF (mm)	RFD	NE1	NE2
0.25	Low	2005	2	Monsoon	31.3	22.5	87	65	50.2	4	0.26	0.29
1.06	High	2006	4	Post monsoon	32.77	19.21	83	41	0	0	2.02	10.24
0.95	High	2006	4	Post monsoon	31.64	15.83	81	39	0	0	1.1	8.96
0.11	Low	2007	4	Post monsoon	36.06	17.47	83	61	0	0	1.98	6.89
0.04	Low	2008	5	Winter	32.37	15.31	87.71	32	0	0	0	4.13
2.16	High	2009	3	Postmonsoon	33.10	19.6	80	42	0	0	0.11	0.17
0.88	High	2010	2	Monsoon	31.4	16.5	88	65	70	2	0.18	0.15
1.2	High	2011	1	Monsoon	30.9	19.5	85	63	0	0	0.38	0
0.36	Low	2013	5	Postmonsoon	30.7	11.4	69	22	0	0	0.76	0.26

PI=*H.armigera* incidence ; MaxT = Maximum temperature; MinT= Minimum temperature; RH1=Morning Relative humidity; RH2=Evening relative humidity;RF= Total rainfall in a week; RFD = Total no. of rainfall days in a week; NE1=No. of Spiders; NE2=No. of coccinellids; Cropstage = 1 (1–5 weeks of the plant); crop stage=2 (square initiation stage); crop stage = 3(flowering and boll formation stage); crop stage = 4 (boll maturity stage); crop stage = 5(boll bursting stage)

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{(ii)}$$

Step 5: The cost value of each bin is calculated as per the formula (iii) and the 'Δ' refers the individual bin of an array.

$$c(\Delta) = \sum_{i=1}^{nb}\frac{2\bar{x} - v}{(n\Delta)^2} \qquad \text{(iii)}$$

Here, $\bar{x}$ and v refers the mean and variance of a bin in an array and n refers the total no. of elements in a bin. The total cost value has been computed as addition of cost of each bin in an array which refers $c(\Delta)$.

Step 6: If c(Δ) <= icost THEN Go to step 9

Step 7: icost = c(Δ)

Step 8: Go to Step 3

Step 9: nob = nb

Step 10: End

The number of bins for an attribute array has been chosen based on the minimum cost value. The total number of bins assigned based on this method in the given data set is for MinT is 5, MaxT is 5, RH1 is 5, and RH2 is 5, RF is 4, NE1 is 5, and NE2 is 5. The program written in Python computer language has been used to implement the optimized binning technique for this data set (Shimazaki and Shinomoto, 2007). The categorical values or labels as A1, A2, A3, A4, A5 have been assigned to these bins to carry out decision rule analysis. For other variables like number of rainfall days in a week (RFD), season and crop stage this method has not been used since they were in the form of nominal values. To the variable RFD, the labels have been assigned as A1, A2, A3, A4, A5, A6, A7 and A8 since the no. of rainy days in a week ranges from 0 to 7 and A1, A2, A3, A4 and A5 have been assigned to the variable crop stage.

ORANGE software

The Excel file of the dataset has been converted into *.CSV file and given as input to ORANGE software ver-

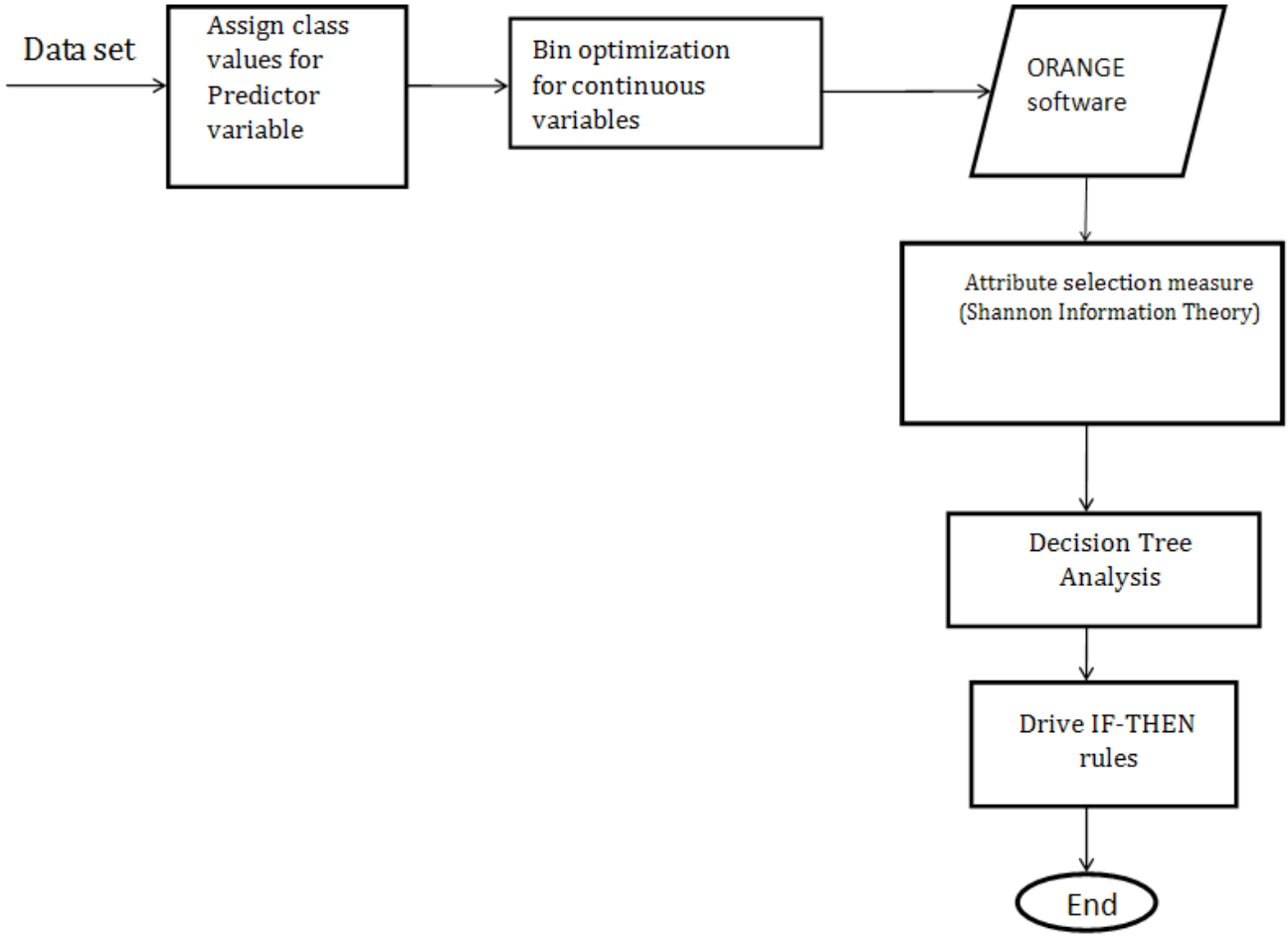

Fig. 1. Data mining process for deriving IF-THEN rules based on bin optimization method.

sion 3:3.3.9 for decision tree analysis. The attribute selection measure Shannon information theory has been used in this analysis to derive IF-THEN rules. The procedure used to derive IF-THEN rules for the pest *H. armigera* incidence by using weather parameters has been given in Fig. 1.

Statistical analysis

The data set was subjected to correlation and regression analysis by using Statistical Package for Social Sciences ver. 17.0 (SPSS, 2008) and represented in Table 2.

RESULTS AND DISCUSSION

Shannon information theory used for attribute selection measure in decision tree analysis and at the time of iteration one for both bin optimization technique and equal width binning method, the gain value based on this theory is given in Table 2 (Gupta, 2006).

The maximum gain value on crop stage is 0.09, followed by MaxT, NE2, MinT, NE1 which shows that among all variables crop stage plays major role on pest incidence. This is confirmed with earlier reports of Leonardo and Miriam (2002) stating that *H. armigera* larvae were confined to succulent plant parts like growing tips, small squares, big squares and bolls hence the crop stage is very important for pest occurrence. Based on the proposed bin optimization technique 31 rules have been derived and based on equal width binning of 5 numbers, 49 rules have been derived. Hence the duplication of rules avoided in bin optimization technique and through this technique the accurate rules can be derived. The

Table 2. Comparison of information gain values, correlation analysis and regression analysis

Attribute/ Variable	Gain value based on bin optimization technique with range of binsize = 2 to 6	Gain value based on equal width binning method with binsize 5	Correlation value (r)	Y = Regression equation
Crop stage	0.09	0.09	-0.011	Y = 2.295-0.146cropstage+0.278season-0.032Maxt-0.037Mint+0.004RF-0.152RFD-0.008RH1+0.012RH2-0.119NE1+0.056NE2 (R^2 = 0.127)
Season	0.03	0.03	0.066	
Maxt	0.07	0.04	-0.086	
Mint	0.03	0.08	-0.043	
RH1	0.02	0.02	0.001	
RH2	0.01	0.01	0.071	
RF	0.01	0	0.124	
RFD	0.02	0.02	-0.047	
NE1	0.03	0.05	-0.091	
NE2	0.07	0.03	0.024	

Table 3. IF-THEN rules for *Helicoverpa armigera* incidence with use of bin optimization technique

S.No	Rule	Predicted Class value of PI	S.No	Rule	Predicted Class value of PI
1.	IF CropStage = 1 AND RH2 = 22 – 34%	Low	17.	IF CropStage = 2 AND RFD = 1	Low
2.	IF CropStage = 1 AND RH2 = 34 – 45%	Low	18.	IF CropStage = 2 AND RFD = 2	High
3.	IF CropStage = 1 AND RH2 = 46 – 57%	Low	19.	IF CropStage = 2 AND RFD = 3	Low
4.	IF CropStage = 1 AND RH2 = 46 – 57% AND RFD = 0	Low	20.	IF CropStage = 2 AND RFD = 4	Low
5.	IF CropStage = 1 AND RH2 = 46 – 57% AND RFD = 1	Low	21.	IF CropStage = 2 AND RFD = 4 AND RH1 = 85-90%	Low
6.	IF CropStage = 1 AND RH2 = 46 – 57% AND RFD = 2	High	22.	IF CropStage = 2 AND RFD = 4 AND RH1 = 90-96%	High
7.	IF CropStage = 1 AND RH2 = 46 – 57 % AND RFD = 0 AND RH1 = 79-84%	Low	23.	IF CropStage = 3	High
8.	CropStage = 1 AND RH2 = 46 – 57% AND RFD = 0 AND RH1 = 85-90%	High	24.	IF CropStage = 4	Low

9.	IF CropStage = 1 AND RH2 = 46 – 57% AND RFD = 1	Low	25.	IF CropStage = 5 AND RH1 = 67-72%	High
10.	IF CropStage = 1 AND RH2 = 46 – 57% AND RFD = 2	High	26.	IF CropStage = 5 AND RH1 = 74 – 78%	Low
11.	IF CropStage = 1 AND RH2 = 59 – 69%	High	27.	IF CropStage = 5 AND RH1 = 79-84%	Low
12.	IF CropStage = 1 AND RH2 = 59 – 69% AND MinT = 15.15 – 18°C	High	28.	IF CropStage = 5 AND RH1 = 79-84% AND NE2 = 0 – 0.78	Low
13.	IF CropStage = 1 AND RH2 = 59 – 69% AND MinT = 18.3 – 21.24°C	High	29.	IF CropStage = 5 AND RH1 = 79-84% AND NE2 = 5.1 – 6.89	Low
14.	IF CropStage = 1 AND RH2 = 59 – 69% AND MinT = 21.3 – 24.31°C	Low	30.	IF CropStage = 5 AND RH1 = 79-84% AND NE2 = 8.28 – 9.62	High
15.	IF CropStage = 1 AND RH2 = 70 – 82%	Low	31.	IF CropStage = 5 AND RH1 = 79-84% AND NE2 = 10.24 – 10.69	High
16.	IF CropStage = 2 AND RFD = 0	High			

important rules based on bin optimization technique have been given in Table 3.

In the time of crop stage 1, when morning Relative Humidity (RH1) is 79-84% and evening Relative Humidity (RH2) is 46-57% and there was no rainy day in that week then the pest incidence was low. But when RH1 was 85-90% and RH2 was 46-57% and RFD was 0 then PI was high. But when RH2 was 59-69% with MinT = 21.3 – 24.31°C, PI was low. When RH2 is 59-69% and MinT was in the range 15.15-18°C or else when MinT was 18.3-21.24°C, then PI was high (Table 3). In the time of crop stage 2, no. of rainy days was 3 in that week, RH1 is 85-90% then PI was low. But, when RH1 was 90-96% PI was high. When crop stage was 3, the pest incidence was high. In the time of crop stage 4, the PI was low. In the time of crop stage 5, when RH1 is between 79-84% and NE2 is 0-0.78 the pest incidence was low. But when RH1 is 79-84 % and NE2 is 10.24 to 10.69, that time the pest incidence was high and this is due to the direct relationship of the pest and the population of Coccinellids.

Correlation analysis reveals that there was no significant coefficient between the PI and other parameters and crop stage, Maxt, Mint and RFD are negatively correlated with pest incidence. The *H.armigera* prediction equation has been derived based on regression analysis with R^2 0.127 (Table 2). In general, correlation analysis helps to find out the significant factors on pest incidence but Shannon information theory is used to derive the gain value based on the probability. But the IF-THEN rules derived from decision tree analysis will be a suitable method for understanding the role of weather parameters on pest incidence since it gives the range values of temperature, humidity, rainfall, etc. Hence, the proposed method can be used for finding the IF-THEN rules for the cause of pest occurrence due to abiotic, biotic and other environmental factors which helps to take up the preventive measures of pest control based on forecasting of the weather in the agricultural field.

REFERENCES

Dhaliwal GS, Arora R. 1996. Integrated pest management: Achievements and Challenges, pp. 308–355. In: Dhaliwal GS, Arora R. (Eds). *Principles of Insect Pest Management*, NATIC, India.

George HJ, Ron K, Karl P. 1994. Irrelevant features and the subset selection problem. In: William W Cohen and Haym Hirsh (Eds.) *Machine Learning: Proceedings of the Eleventh International Conference.* 121-129, Morgan Kaufmann Publishers, San Francisco, CA.

Gupta GK. 2006. Classification. In: Introduction to Data Mining with Case Studies, Prentice-Hall of India, 106–136. https://doi.org/10.1016/B978-044451636-7/50013-9

Leonardo T, Miriam EP. 2002. The distribution and movement of cotton bollworm, *Helicoverpa armigera* Hübner (Lepidoptera: Noctuidae) larvae on cotton. *Philippine J Sci,* **131**: 91–98.

Pratheepa M, Meena K, Subramaniam KR, Venugopalan R, Bheemanna H. 2011. A decision tree analysis for predicting the occurrence of the pest, *Helicoverpa armigera* and its natural enemies on cotton based on economic threshold level. *Curr Sci.* **100**(2): 238–246.

Shimazaki H, Shinomoto S. 2007. A method of selecting the binsize of a Time Histogram. *Neural Comput.***19**(6): 1503–1527.

SPSS V 17.0. 2008. *Statistical Package for Social Sciences.* SPSS Inc. Illinois, Chicago,USA.

Sotiris K, Dimitris K. 2006. Discretization techniques: A recent survey. *GESTS International Trans Comput. Sci Engineering.* **32**(1): 47–58.

Zhao H, Ram S. 2004. Constrained cascade generalization of decision trees. *IEEE Trans Knowledge Data Engineering*. **16**(6): 727–739. Available from: https://dl.acm.org/citation.cfm?id=1437601 https://doi.org/10.1109/TKDE.2004.3