Journal of Biological Control, 32(1): 68-73, 2018, DOI: 10.18311/jbc/2018/16309

Research Article

A Bayesian classification approach for predicting *Gesonia gemma* Swinhoe population on soybean crop in relation to abiotic factors based on economic threshold level

J. CRUZ ANTONY[1,2] and M. PRATHEEPA[1*]

[1]*Division of Genomic Resources, ICAR-National Bureau of Agricultural Insect Resources, Bengaluru – 560024, Karnataka, India*

[2]*Department of Computer Science, Jain University, Bengaluru – 560027, Karnataka, India*

[*]*Corresponding author E-mail:* mpratheepa@gmail.com

ABSTRACT: Predicting of insect pest population with accuracy and speed when given large data set will make a major contribution to the success of integrated pest management. Naïve Bayesian classification has been proposed for predicting the insect pest *Gesonia gemma* Swinhoe on soybean crop. The Naïve Bayesian classifier works based on Bayes' theorem and can predict class probabilities that a given tuple from the dataset belongs to a particular class. The dataset includes abiotic factors as features along with the class feature (pest incidence) are separated as training data and testing data, then the model was built on the training set by finding the probability for each of its features in relation with the class feature. The Naïve Bayesian classification from the trained model, best fits the testing data with 90% accuracy, thus the proposed approach can be very useful in predicting the pest *G. gemma* on soybean crop.

KEY WORDS: Abiotic, Bayesian classification, *Gesonia gemma*, Naïve population dynamics, soybean

(Article chronicle: Received: 14-07-2017; Revised: 26-09-2017; Accepted: 30-09-2017)

INTRODUCTION

Soybean [*Glycine max* (L.) Merrill] is one of the most important oilseeds crop cultivated around the world. According to the latest report by Foreign Agricultural Service (USDA), the production constitutes around 60% of the total world production of oilseeds and figures around 325 million metric tons. India holds the fifth position next to China among largest producers of soybean in the world. Soybean is mainly cultivated in the states of Madhya Pradesh, Maharashtra, Rajasthan, Karnataka, Andhra Pradesh and Chattisgarh (Agarwal *et al*., 2013). Soybean in India is infested by about 20 major insect species, of which semiloopers are important. A species complex comprising of *Gesonia gemma* Swinhoe, *Chrysodeixis acuta* (Wlk.), *Diachrysia orichalcea* (Fabricius) and *Mocis undata* Fabricius differ in colour, shape, and size are generally noticed on soybean in Maharashtra (Sharma, 2011). The *Gesonia gemma* Swinhoe (1885), the grey semilooper, causes defoliation to the crop, which imposes significant loss in grain by reducing its weight (Yadav *et al*., 2014). The *G. gemma* population on soybean crop from Nagpur district of Maharashtra was assessed from the year 2009 to 2013, for studying its population dynamics. According to (Southwood, 1977), the reason for pest population dynamics is due to abiotic factors rather than biotic factors. Hence, the prediction of pest based on abiotic factors is necessary. Several techniques are available for pest prediction but there is a gap in understanding the factors responsible for pest flare up. In view of this, the data mining technique Naïve Bayesian (NB) classification has been proposed for predicting the population of *G. gemma* on soybean crop based on trained model, as it is an important part of evaluating the model and also it can minimize the effects of data discrepancy and better understand the characteristics of the model. The optimum rules were also learned from the NB model, which makes it easy to understand the population dynamics of *G. gemma* on soybean. The NB classifier is considered simplest and computationally most efficient (Han and Kamber, 2006). The NB classification is a widely used framework for classification and it works with the Bayes' theorem (Good, 1965). The advantage of NB classification is mainly on attribute inde-

pendence assumption (Zaidi *et al*., 2013), because of which it suits for finding the role of each of the factors related with pest incidence.

METHODOLOGY

Dataset

The dataset was obtained from a collaborative project on crop pest surveillance implemented by the State Department of Agriculture, Maharashtra. Field pest scouting data from Nagpur district of Maharashtra along with its abiotic factors were used in the model for analysis. The pest incidence with mean values of previous two weeks weather data for parameters *viz*., maximum temperature (MaxT), minimum temperature (MinT), Relative Humidity (RH) and total rainfall in mm (RF) of two weeks along with moisture adequacy index (MAI %), soil moisture index (SMI %) and total number of rainy days (RFD) were taken for analysis. The sample dataset is given in Table 1.

Data pre-processing

The continuous numerical features MaxT, MinT, RH, MAI, SMI, and RF are subjected to data discretization and Max-diff discretization technique was used to transform these continuous numerical features into categorical counterparts (Antony and Pratheepa, 2016). This method calculates the maximum difference between each value of its respective feature and groups according to the number of bins required for the features. Five bins have been created for each feature and bin labels were named as A1, A2, A3, A4 and A5 which have been chosen arbitrarily. The pest population count has been classified as high, medium and low based upon the Economic Threshold Level (ETL) and this was defined using the feature crop stage (pre-flowering and post-flowering) and standard week. The standard week from 27 to 33 is considered as the pre-flowering stage of the crop, in which the pest incidence (per meter row length) between 0 and 3.0 was classified as low, 3.1 to 6.0, was classified as Medium and greater than 6.0 as high. The standard week from 34 to 41 is considered as the post-flowering stage of the crop, in which the pest incidence between 0 and 2.5 was classified as low, 2.6 to 5.0 was classified as medium and greater than 5.0 as high. The sample transformed data have been given in Table 2 and the bin range values for each feature has been given in Table 3.

Training and testing phase

Generally, the NB classification model goes through two phases. The first phase is the learning phase where the classification model is built from the training dataset using probabilistic approach. The second is the application phase where the system applies the learned classification model to infer or predict the classes of the new test dataset. In this regard, the dataset has to be divided into training and test data and 80% from the dataset has been assigned for training dataset and remaining 20% for the test set. The dataset comprises of 613 tuples of which 30 tuples were classified as high, 42 tuples were classified as medium and majority 541 tuples were as low. It is clearly understandable that the dataset is imbalanced; a perfect ratio has to be measured for choosing the data for training and test datasets. The ratio

Table 1. Sample tuples from the dataset

PI	SW	CS	MaxT (°C)	MinT (°C)	RH (%)	MAI (%)	SMI (%)	RF (mm)	RFD
0.0	27	1	30.00	21.50	95.00	100	72	69.0	3
0.0	28	1	27.50	22.50	95.00	100	82	43.0	4
2.8	29	1	27.00	23.00	89.58	100	100	89.9	4
22.2	30	1	28.90	23.67	89.33	100	100	25.3	2
53.0	31	2	33.20	24.00	76.25	95	89	3.0	1
9.7	32	2	30.60	23.65	80.50	83	72	4.5	1
1.2	33	3	32.19	24.25	81.25	75	60	0.0	0
2.8	34	3	31.12	25.50	88.75	100	99	202.6	4
3.7	35	4	28.85	24.33	90.50	100	100	111.8	3
2.6	36	4	27.69	23.94	87.75	100	100	27.0	1
0.9	37	4	29.36	24.41	82.25	98	84	24.6	3

PI - pest incidence; SW - standard week; CS - crop stage (1 - vegetative stage I; 2 - vegetative stage II; 3 - flowering stage; 4 - fruiting stage); MaxT - maximum temperature; MinT - minimum temperature; RH - relative humidity; MAI - moisture adequacy index; SMI - soil moisture index; RF - rainfall in mm; RFD - number of rainfall days in a week.

was calculated based upon the number of tuples classified as high, medium and low (i.e., 30, 42 and 541 out of total 613), and thus the ratio 1:1.5:18 was obtained. Then using this ratio, 80% (490 tuples) were chosen randomly and assigned to the training dataset and remaining 20% (123 tuples) for test dataset.

Naïve Bayesian classification model

Let *D* be a training set of tuples with their associated class labels. Each tuple is represented by an n-dimensional feature vector, $X = (X1, X2, \ldots Xn)$ with 'n' features (MaxT, MinT, RH, MAI, SMI, RF and RFD) depicting the data discretization technique made on the tuple with discretization values $A1, A2, \ldots Am$, with class values C_i (where i = 1 to 3) which represents three classes namely high, medium and low. Bayesian classifiers are statistical classifier which predicts the class of unseen instance by conjunction of the feature values $X = a1^{a2^{\cdots an}}$ (Duda and Hart, 1973). The predicted class is based on highest posterior probability $P\left(\frac{C_i}{X}\right)$ and is defined as follows:

$$P\left({}^{C_i}/_{X}\right) = \frac{P(C_i)\, P\left({}^{X}/_{C_i}\right)}{P(X)} \quad \text{Eq. (1)}$$

where $P\left({}^{X}/_{C_i}\right) = \prod_{k=1}^{n} P\left({}^{x_k}/_{C_i}\right)$

$$= P\left({}^{x_1}/_{C_i}\right) \times P\left({}^{x_2}/_{C_i}\right) \times \cdots \ldots \times P\left({}^{x_n}/_{C_i}\right) \quad \text{Eq. (2)}$$

and $P(X)$ is constant across all the classes and can be neglected.

$$P(C_i) = \frac{[\![n(C]\!]_{i,D})}{n(D)}$$ where $[\![n(C]\!]_{i,D})$ denotes total number of records belongs to the class C_i and $n(D)$ denotes total number of tuples in the dataset.

Assume a tuple i.e., a feature vector *X* = {A3, A3, A3, A5, A5, A1, A1} from the dataset, where the class feature C_i has to be predicted. Based on Equation (1), the values for the terms $[\![P(C]\!]_i)$ and $P\left({}^{X}/_{C_i}\right)$ for each of the class values has to be calculated. First, $[\![P(C]\!]_i)$ was calculated for all three class values i.e. , $[\![P(C]\!]_{Medium})$ $[\![P(C]\!]_{Medium})$ and $[\![P(C]\!]_{Low})$. The total number of tuples from training dataset is 490, out of which 24 records with class High, 33 records with class Medium and remaining 433 records with class Low. The probabilities of these class values are shown below,

Table 2. Sample transformed tuples from the dataset

P.I	MaxT (°C)	MinT (°C)	RH (%)	MAI (%)	SMI (%)	RF (mm)	RFD
Low	A3	A3	A5	A5	A3	A1	A2
Low	A1	A2	A4	A5	A4	A1	A3
Low	A1	A2	A5	A5	A5	A1	A3
High	A3	A3	A5	A5	A5	A1	A2
High	A3	A3	A2	A5	A5	A1	A1
Medium	A3	A3	A4	A4	A4	A1	A1
Low	A3	A3	A4	A4	A3	A1	A1
low	A3	A3	A5	A5	A5	A2	A5
Medium	A3	A3	A5	A5	A5	A1	A4
Low	A3	A2	A4	A5	A5	A1	A1
Low	A3	A3	A4	A5	A4	A!	A2

Table 3. Bin range values of the features

Abiotic features	A1	A2	A3	A4	A5	A6	A7	A8
MaxT	27.25 - 28.1	29.5	29.9 - 33.7	35.29	35.83	-	-	-
MinT	19.71	22.29 - 22.4	22.86 - 24.83	25.04	25.5	-	-	-
RH	76.67 - 80.83	82.6 - 84	85	86.14 - 88.4	89.3 - 96	-	-	-
MAI	86	88 - 89	92	94 - 97	99 - 100	-	-	-
SMI	16	67	72	78 - 85	89 - 100	-	-	-
RF	0 - 224	234 - 242.4	255.5	284.5	367.6	-	-	-
RFD	0	1	2	3	4	5	6	7

$[\![P(C)]\!]_{High}$ = (24/490) = 0.048979592
$[\![P(C)]\!]_{Medium}$ = (33/490) = 0.0673469
$[\![P(C)]\!]_{Low}$ = (433/490) = 0.8836735

The conditional probability $P\left(X/_{C_i}\right)$ has been calculated based on Equation (2) and the probability calculation for $P\left(X/_{C_i}\right)$ of the class values are shown below,

Computing $P\left(X/_{C_i}\right)$ for class *High*

$[\![P(MaxT = A3 \mid C]\!]_{\downarrow High}$ = (23/24) = 0.96
$[\![P(MinT = A3 \mid C]\!]_{\downarrow High}$ = (24/24) = 1
$[\![P(RH = A3 \mid C]\!]_{\downarrow High}$ = 0
$[\![P(MAI = A5 \mid C]\!]_{\downarrow High}$ = (17/24) = 0.71
$[\![P(SMI = A5 \mid C]\!]_{\downarrow High}$ = (22/24) = 0.92
$[\![P(RF = A1 \mid C]\!]_{\downarrow High}$ = (24/24) = 1
$[\![P(RFD = A1 \mid C]\!]_{\downarrow High}$ = (5/24) = 0.21

Substituting in Equation (2),

$[\![P(X \mid C]\!]_{\downarrow High}$ = (0.96 x 1 x 0 x 0.71 x 0.92 x 1 x 0.21) = 0

Similarly, computing $P\left(X/_{C_i}\right)$ for class *Medium and Low*

$[\![P(X \mid C]\!]_{\downarrow Medium}$ = (0.91 x 1 x 0.03 x 0.79 x 0.85 x 1 x 0.09) = 0.0016498755

$[\![P(X \mid C]\!]_{\downarrow Low}$ = (0.85 x 0.85 x 0.05 x 0.82 x 0.81 x 0.99 x 0.12) = 0.00285051

Laplacian correction

Consider the value for $[\![P(X \mid C]\!]_{\downarrow High}$, since there were no training records for $[\![P(RH = A3 \mid C]\!]_{\downarrow High}$, the final value for $[\![P(X \mid C]\!]_{\downarrow High}$ ended up with zero. Plugging this zero value into the Equation (1) will return a zero probability for $P\left(C_{High}/_X\right)$, even though, without the zero probability, it may have ended up with highest posterior probability. Hence to avoid this zero probability, Laplacian correction or Laplace estimator technique was used by adding one to each count of training dataset *D* which would make a minor negligible difference and thus avoiding the zero probability value. So reconsidering the value for $[\![P(X \mid C]\!]_{\downarrow High}$,

$[\![P(X \mid C]\!]_{\downarrow High} = \frac{24}{31}$ x $\frac{25}{31}$ x $\frac{1}{31}$ x $\frac{18}{31}$ x $\frac{23}{31}$ x $\frac{25}{31}$ x $\frac{6}{31}$
= (0.77 x 0.81 x 0.03 x 0.58 x 0.74 x 0.81 x 0.19)
= 0.00123593

It has been observed that one count was added to each of the feature vector *X* (i.e., the numerator) then the count for the class C_{High} (i.e. the denominator) increases to 7, which add up from 24 to 31. In this way closer value to the actual value of $[\![P(X \mid C]\!]_{\downarrow High}$ was obtained to avoid zero probability.

After completion of finding conditional probability $P\left(X/_{C_i}\right)$ and the probability of class values $[\![P(C)]\!]_i$, both these values are applied in Equation (1) for obtaining posterior probability of all three class values,

$[\![P(C]\!]_{\downarrow High} \mid X) = [\![P(C)]\!]_{High}$ x $[\![P(X \mid C]\!]_{\downarrow High}$
= 0.048979592 x 0.00123593 = 0.00006054
$[\![P(C]\!]_{\downarrow Medium} \mid X) = [\![P(C)]\!]_{Medium}$ x $[\![P(X \mid C]\!]_{\downarrow Medium}$
= 0.0673469 x 0.0016498755 = 0.0001111
$[\![P(C]\!]_{\downarrow Low} \mid X) = [\![P(C)]\!]_{Low}$ x $[\![P(X \mid C]\!]_{\downarrow Low}$
= 0.8836735 x 0.00285051 = 0.0025189

The probabilistic values of $[\![P(C]\!]_{\downarrow High} \mid X)$, , and $[\![P(C]\!]_{\downarrow Low} \mid X)$ have been compared and since $[\![P(C]\!]_{\downarrow Low} \mid X)$ was having highest probability value, the feature vector *X* = {A3, A3, A3, A5, A5, A1, A1} was classified as Low.

The above computation process was repeated for every tuple in the test dataset, in order to predict its classification.

RESULTS AND DISCUSSION

The Bayesian rules have been derived by applying a threshold value to all tuples from the test set where it has been classified correctly by using the well-built NB model. The threshold value is fixed as "0.5", so the highest posterior probability value for a tuple should be greater than 0.5 and those tuples were extracted as rules. Some of the important Bayesian rules derived from the model have been shown in Table 4.

The Bayesian rules derived from NB model reveals that in Nagpur district, when maximum temperature was between 29.9˚C and 33.7˚C and minimum temperature was between 22.86˚C and 24.83˚C with relative humidity between 89.29% and 96%, moisture adequacy index between 99% and 100% and soil moisture index between 89% and 100%, rainfall between 1 and 242.4 (mm) and no. of rainy days between 3 and 5, then the pest incidence was low. Similarly, when maximum temperature was between 29.9˚C and 33.7˚C and minimum temperature was between 22.86˚C and 24.83˚C with relative humidity between 76.67% and 84%, moisture adequacy in-

Table 4. Bayesian rules derived from the model

S.No	Learned Bayesian classification rules
1	(MaxT = 29.9 - 33.7) ^ (MinT = 22.86 - 24.83) ^ (RH = 89.29 - 96) ^ (MAI = 99 - 100) ^ (SMI = 89 - 100) ^ (RF = 234 - 242.4) ^ (RFD = 5) -> **(PI = Low)**
2	(MaxT = 29.9 - 33.7) ^ (MinT = 22.86 - 24.83) ^ (RH = 89.29 - 96) ^ (MAI = 99 - 100) ^ (SMI = 89 - 100) ^ (RF = 1 - 224) ^ (RFD = 6) -> **(PI = Low)**
3	(MaxT = 29.5) ^ (MinT = 22.86 - 24.83) ^ (RH = 89.29 - 96) ^ (MAI = 99 - 100) ^ (SMI = 89 - 100) ^ (RF = 1 - 224) ^ (RFD = 3) -> **(PI = Low)**
4	(MaxT = 29.9 - 33.7) ^ (MinT = 22.86 - 24.83) ^ (RH = 76.67 - 80.83) ^ (MAI = 94 - 97) ^ (SMI = 89 - 100) ^ (RF = 0) ^ (RFD = 0) -> **(PI = High)**
5	(MaxT = 29.9 - 33.7) ^ (MinT = 22.86 - 24.83) ^ (RH = 82.6 - 84) ^ (MAI = 94 - 97) ^ (SMI = 89 - 100) ^ (RF = 1 - 224) ^ (RFD = 1) -> **(PI = High)**
6	(MaxT = 29.9 - 33.7) ^ (MinT = 22.86 - 24.83) ^ (RH = 86.14 - 88.4) ^ (MAI = 99 - 100) ^ (SMI = 89 - 100) ^ (RF = 1 - 224) ^ (RFD = 2) -> **(PI = Medium)**

Table 5. Confusion matrix derived from test set

	Actual class				
		High	Medium	Low	Predicted overall
Predicted class	High	2	0	0	2
	Medium	0	1	0	1
	Low	4	8	108	120
	Actual overall	6	9	108	123

dex between 94% and 97% and soil moisture index between 89% and 100%, rainfall between 0 and 1 (mm) and no. of rainy days between 0 and 1, then the pest incidence was high. And when maximum temperature was between 29.9°C and 33.7°C and minimum temperature was between 22.86°C and 24.83°C with relative humidity between 86.14% and 88.4%, moisture adequacy index between 99% and 100% and soil moisture index between 89% and 100%, rainfall between 1 and 224 (mm) and no. of rainy days between 1 and 2, then the pest incidence was medium.

To determine the performance and prediction accuracy of the NB classifier, a confusion matrix was computed for test dataset and has been shown in Table 5.

It has been clearly observed that out of 123 test set tuples, 111 tuples have been predicted exactly and the overall accuracy of the model is 90.244% and the Cohen's Kappa coefficient value is computed as 0.313 which is a fair agreement of the model (Kappa statistic agreement: values ≤ 0 as indicating no agreement and 0.01-0.20 as none to slight, 0.21-0.40 as fair, 0.41- 0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1.00 as almost perfect agreement).

CONCLUSION

The Naïve Bayesian classification model was proposed for predicting the *Gesonia gemma* population related to abiotic factors in Nagpur district of Maharashtra. The Bayesian rules were also been learned from the model, which has exhibited clearly that the higher temperature and moderate percentage of humidity with very low rainfall were the favorites of the pest population. Similarly, when the temperature was moderate with high humidity and more rainfall days, then the population seems to be very less on the crop. Thus, the Bayesian rules learned from the model makes it easy to study the population dynamics of *Gesonia gemma* occurring on soybean, which will help the farmers to make necessary preparations in time to control pest and also provide a major contribution to the success of integrated pest management for soybean.

ACKNOWLEDGEMENT

Authors are thankful to the Director, ICAR-NBAIR, Bengaluru for providing facilities to carryout the research work. Authors are also thankful to the editor of JBC for providing valuable comments to improve the manuscript. The dataset used in the study was obtained from CROPSAP programme of Maharashtra, Commissionerate of Agriculture, Government of Maharashtra.

REFERENCES

Agarwal DK, Billore SD, Sharma AN, Dupare BU, Srivastava SK. 2013. Soybean: Introduction, improvement, and utilization in india-problems and prospects. *Agric Res.* **2**(4): 293–300. https://doi.org/10.1007/s40003-013-0088-0

Antony JC, Pratheepa M. 2017. Study of population dynamics of soybean semi-looper *Gesonia gemma* Swinhoe

by using rule induction model in Maharashtra, India. *Legume Res.* **40**(2): 369–373.

Duda RO, Hart PE. 1973. *Pattern classification and scene analysis.* John Wiley, NY.

Good IJ. 1965. *The estimation of probabilities: An essay on modern Bayesian methods.* M.I.T. Press, Cambridge, MA.

Han J, Kamber M. 2006. *Data mining: Concepts and techniques.* Morgan Kaufmann, San Francisco, CA.

Sharma AN. 2011. Insect pests of soybean: Present management strategies and future thrusts. 3rd Congress on Insect Science - Pest Management for Food Security and Environment Health, organized by Indian Society for the Advancement of Insect Science, Ludhiana, India, 18th-20th Apr 2011.

Southwood TRE. 1977. The relevance of population dynamics theory to pest status. In: Cherret JM and Sagar GR (Eds.) *Origins of Pest, Parasite, Disease and Weed Problems.* Blackwell Scientific Publications, Oxford, UK.

Zaidi NA, Cerquides J, Carman MJ, Webb GI. 2013. Alleviating Naïve Bayes attribute independence assumption by attribute weighting. *J Mach Learn Res.* **14**: 1947–1988.

Yadav SS, Nayak MK, Srivastava AK, Gupta MP, Tomar DS. 2014. Population dynamics of insect defoliator of soybean and correlation with weather parameters. *Ann Plant Protect Sci.* **22**: 208–209.