An Approach Towards Development of a Stem Borer Population Prediction Model Using R Programming

Sudipta Paul[1(✉)], Sourav Banerjee[2], and Utpal Biswas[1]

[1] Department of Computer Science and Engineering, University of Kalyani, Kalyani 741235, India
sudiptap48@gmail.com, utpal0lin@yahoo.com
[2] Department of Computer Science and Engineering, Kalyani Government Engineering College, Kalyani 741235, India
contact200683@gmail.com

Abstract. The rice is a major crop of India. It is the staple food of the eastern and southern parts of this country. The total yield of rice can be in a massive loss if it is affected by pests. The stem borer pest creates a lot of trouble. It affects the production of rice. As the control procedure with pesticide is not much effective on this pest, therefore, a forecasting model can play a major role in taking preventive measure. The objective of this research is to forecast the population occurrence of stem borer pest in the paddy. This paper highlights the improvement of the performance of backpropagation artificial neural network (BP-ANN) model using principal component analysis (PCA) to develop a prediction model by minimizing the error. The Convolution of data is proposed here instead of PCA to enhance the reduction of the dimensions of data which eventually results in less error in prediction.

Keyword: Backpropagation artificial neural network (BP-ANN) · Principal component analysis (PCA) · Stem borer · Convolution

1 Introduction

The stem borer (*Scirpophaga Incertulas)* is widespread in all world, especially in Asia. The stem borer is a major pest of paddy in India. This pest is endemic and is distributed in most parts of India. Insecticides are not a good option to control this kind of pest because of its monophagous nature and peculiar boring habit. Therefore, a forecast is highly needed to take early measure to prevent the appearance of this pest. Here we did not get any Indian statistics data regarding the appearance of this pest through a period of definite times of years. This is why we are using here the database from the work Yang et al. [1], in 2009. At first, we implemented according to their work. After that, we enhanced it slightly to get a better result in prediction with less error rate using convolution in place of principal component analysis. Our proposed work results better than the existing works with respect to the error rate and dimension reducing of raw, uncleaned data for prediction of the population occurrence of stem borer in paddy. Artificial Neural Network (ANN) [15] is used in many fields of science because of its

J. K. Mandal et al. (Eds.): CICBA 2018, CCIS 1031, pp. 145–154, 2019.
https://doi.org/10.1007/978-981-13-8581-0_12

different characteristics which are very special in its own way, such as parallel calculations in a large-scale, storage of information which is distributed in nature, adaptability with its own capability, organization in its own capability and high fault-tolerant properties, and also for nonlinear relations it has a very good simulation of fitness. Thus ANN has its widespread use in such areas as pattern recognition, handling of knowledge, nonlinear transformation, the technology regarding remote sensing, research on the robot, and projects on biomedicine according to the work of Yang et al. [1]. Among a lot of various Artificial Neural Network algorithm, the Back Propagation (BP) network is used mostly. Generally, the transmission functions used here are nonlinear in nature. The common transfer functions are the logarithm S type (logsig) function and the hyperbolic tangent S type (tansig) function. In a BP network, the linking between neurons works in a front feedback neural network manner and the learning method works in a supervision study manner. A Convolutional Neural Network (CNN) [8] has recently been popularised for its success in classification in several kinds of problems (e.g. image recognition [13] or time series classification [14]). This network consists of a sequence of convolution layers. The output in the layers is passed only to local regions of theinput. This is achieved by doing the dot product between the input and the filter sliding at each data-point. This whole process of calculation is called convolution. This convolution process allowsthe CNN filters to know and acknowledgea specific pattern in the given data-sets.

According to [8], the convolution of two one-dimensional signals f and g, which is discrete in nature, is written as $f * g$ and defined as:

$$(\mathrm{f} * \mathrm{g})(\mathrm{i}) = \sum_{\mathrm{j}=-\infty}^{\infty} \mathrm{f}(\mathrm{j})\mathrm{g}(\mathrm{i}-\mathrm{j}) \tag{1}$$

Here, definition of the convolution states that, the nonexistent data-samples in the input may have values of zero, it is often described as zero padding. Therefore it is evident that the computing of the products take place only at the data-points where samples exist in both filter and data set. A point is to be noted here that the process of convolution is commutative, i.e. $(f * g) = (g * f)$. If the data-pointsare finite, the infinite convolution is permitted to be truncated. In other words, suppose f = [f(0), …, f(N − 1)] and g = [g(0), …, g(M − 1)], the convolution is,

$$(\mathrm{f} * \mathrm{g})(\mathrm{i}) = \sum_{\mathrm{j}=0}^{\mathrm{m}-1} \mathrm{f}(\mathrm{j})\mathrm{g}(\mathrm{i}-\mathrm{j}) \tag{2}$$

The current models for the forecasting of population dynamics of stem borer inpaddy have three major shortcomings with some benefits [1, 10]: the ability of data-fitting is insufficient, the generality of the whole model is weak and the error rate between reality and the predicted result is too large. This paper develops a new model of prediction using BP ANN with Convolution of data matrix using a suitable filter, to find a non-linear relation between the population of stem borer and the main meteorological factors, then using ANN build a prediction model for the population occurrence of paddy stem borer.

2 Related Works/Literature Review

After a thorough background work some of the valuable works on the field are the following:

In 2009, Yang et al. [1] has published "A prediction model for population occurrence of paddy stem borer (Scirpophaga Incertulas), based on Back Propagation Artificial Neural Network and Principal Components Analysis". In this paper they have done a survey at the Plant Protection Station of JianShui County, Yunnan and associated meteorological data were obtained from the JianShui County Meteorologic Observatory in China. In this paper, they have applied PCA and BP-ANN methods to analyze the aforesaid obtained data on population occurrence to find out a non-linear relation between the stem borer population and the meteorological factors to build a prediction model which have a good accuracy level of prediction. We took it as the base prediction model and then enhanced this model on the basis of their database in our project. Rad et al. [2] in 2014, also have made another prediction model using Artificial Neural Network. The estimation of the non-linear relation between the data-point in the data set can be very helpful for calculating the amount of variance of a particular data-point with the comparison to other data-point in a dataset. This paper aimed to calculate the variance of different agronomic and phenologic factors on the total mass of melon fruit produced. Sun et al. [3] in 2015 have also compared some existing prediction models in their paper regarding rice strip virus by using three different models: stepwise regression, back propagation neural network, and support vector machines. Günther and Fritsch [4] have described the way the package "neuralnet" work, its pros and cons in their paper regarding its use in R statistics software. In 2009, Smith et al. [5], has coined an ANN based model for year around temperature prediction. They have explored various applications of ANNs for the prediction of temperature during the entire year. Their ANNs were developed using detailed data collected by the Georgia Automated Environmental Monitoring Network (AEMN). The ANNs were able to give predictions with a mean absolute error (MAE) which was less during the winter months than the MAE of the previously developed winter-specific models. In 2015, Sengar and Kalpana [6] has shed some light regarding the climate or atmospheric conditions changing effects on paddyyielding. Atmospheric conditions are significant factors in the distribution, yielding, and security of food. This book discusses in global detail, with special reference to India about the recommendations for achieving climate-smart agriculture. Javad et al. [7], in 2016, shed some light on the determination of a model on a prediction of soil cation exchange capacity. This project shed some light on the comparison of multiple linear regression, multiple non-linear regression, adaptive neuro-fuzzy inference system and artificial neural network including feed-forward back propagation (FFBP) model to calculate the soil cation exchange capacity in Guilan province, northern Iran. In 2017, Borovykh et al. [9] present a method for conditional time series forecasting based on the CNN architecture. The proposed model contains stacks of convolutions between filters and the datasets which allow an access of a broad range of data-points when predicting; multiple filters are applied here in parallel to separate time series datasets and allow the

fast processing of data and the utilization of the correlation structure between the multivariate time series.

3 Proposed Work

Here, we are using the database used by Yang et al. [1], in 2009. According to their survey, the factors that are influential for the population occurrence of stem borer are sixand stated in Tables 1 and 2. Therefore, the six factors from Table 1 and another batch of six factors from Table 2 give all total 12 influencing factors for the proposed model.

Table 1. The population occurrence of paddy stem borer in the 1st generation and its relative meteorological factors' data in March from 2000 to 2008

Year	March						The population occurrence in 1st generation
	Average temperature (c)	Maximum temperature (c)	Minimum temperature (c)	Rainfall (mm)	Potential evaporation (mm)	Relative humidity (%)	
2000	17.3	28.7	6.5	39.2	203.6	64	92
2001	18.5	29.7	10.3	25.7	222.2	63	132
2002	18.9	30.3	6	26.8	267	58	30
2003	17.9	28.2	7.8	32.5	250.9	61	16
2004	19.6	31.6	6	0.2	250.9	57	207
2005	16.6	30	2.5	44.5	201.8	64	201
2006	19.5	30.3	6.2	1.3	226.7	53	128
2007	20.1	31.6	7.8	0.5	272.8	48	191
2008	17.2	28.9	3.4	28.5	161.8	64	60

Table 2. The population occurrence of paddy stem borer in the 2nd generation and its relative meteorological factors' data in April from 2000 to 2008

Year	April						The population occurrence in 2nd generation
	Average temperature (c)	Maximum temperature (c)	Minimum temperature (c)	Rainfall (mm)	Potential evaporation (mm)	Relative humidity (%)	
2000	21.7	31	11.7	16.6	259.8	62	92
2001	23.1	32.2	12.7	18.4	321.8	50	132
2002	22.2	31.7	11.3	39.1	308.5	56	30
2003	22.9	32.5	13.1	7	326.3	53	16
2004	20.1	31.1	11.4	104	213.7	67	207
2005	21.6	32.8	8.8	37.9	258	60	201
2006	22.3	30.3	12.4	85.7	241.2	54	128
2007	19.2	30.9	9.2	108.6	179.1	66	191
2008	22	33	12.4	38.6	251.9	59	60

Table 3. The proposed steps

Steps	Description
Model type	A three-layer BP network because of its ability to approximate any function
Transfer function	$\frac{1}{1+e^{-x}}$
Number of neurons in the hidden layer	10 neurons
Input layer	Considering the convoluted data set that influence the paddy stem borer population occurrence, take m elements as the input which have the least value in its columns. Therefore there are m neurons in the input layer of BP ANN
Output layer	One Neuron. This Neuron will state the occurrence of the population of the stem borer

Step 1: **Normalizing the dimension of the datasets using convolution of data matrix**
As there are all total 12 parameters, to draw a neural network with this much parameters is a chaotic work. Therefore it has been taken account the process of convolution of data matrix which will eventually use its power per parameter per input (PPPPI) to extract features by taking advantage of the structural information of the datasets. Also it will reduce the dimension of the datasets too. In the process of taking convolution at first the dataset is padding with extra "zeros" in the borders of the dataset. Therefore 2 extra rows and 2 extra columns of data is introduced here. It will help not to lose any essential data variation from the original dataset. After that a kernel matrix of dimension 3×3 is taken where all the diagonal values are 1. Taking a kernel matrix is a tricky work, because there is no hard and fast rule. A lot of kernels have been tried and only after the trial and error method the aforesaid kernel gave the best result.
As the main aim is to reduce the dimension of the datamatrix efficiently the convolution Quantile() function of the imagine package in the R statistical software is used here. Here the quantile function will divide each of the data set of values with a variance, which further divide a frequency distribution of equal groups, each contained the same fraction of the total frequency of the data sets.

Step 2: **Backpropagation Neural Network with the convoluted data matrix as input neuron**
From [1] (Chen et al. [11], Yang et al. [1]) it is known that the three-layer ANN model has the ability to simulate any nonlinear system of equations. Thus, any continuous time series or map type data sets can be easily modelled by a three-layer ANN. The proposed steps to build a BP ANN model for predicting the population occurrence of stem borer in paddy is stated in the following Table.

4 Experimental Results and Comparisons

According to the proposed model in Sect. 3 at first we normalize the dimension of the datamatrix with the given method of convolution, then we simply follow the steps of the back propagation Neural Network to implement the model. Following are the discussion of the results of the proposed work.

4.1 Result of the Proposed Method

Step 1: **Normalize the dimension of the datasets using convolution of data matrix**

Table 4. Normalized datasets using convolution of data matrix with less dimension

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
Y1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Y2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Y3	NA	NA	6.0	10.3	25.7	22.9	22.2	13.1	7.0	12.7	18.4	16
Y4	NA	NA	0.2	6.0	26.8	20.1	22.9	11.4	13.1	7.0	39.1	53
Y5	NA	NA	6.0	0.2	32.5	21.6	20.1	8.8	11.4	13.1	7.0	67
Y6	NA	NA	1.3	6.0	0.2	22.3	21.6	12.4	8.8	11.4	54.0	60
Y7	NA	NA	0.5	1.3	44.5	19.2	22.3	9.2	12.4	8.8	37.9	54
Y8	NA	NA	7.8	0.5	1.3	22.0	19.2	12.4	9.2	12.4	59.0	60
Y9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Here, it is being seen that the dataset has been reduced quite an amount in its dimension. Here the rows and columns of "NA" indicates the extra padded zeros which have been introduced here to help not to lose any data variation. Also it is now easier to handle them as inputs in the backpropagation neural network because of its less dimension. Here E1 to E12 are the columns of data elements from which the columns with least range of variation will be taken as input in the ANN. Also Y1 to Y9 indicate the effective factors in the dataset (Table 4).

Here the kernel matrix is following (Table 5):

The quantile function in R programming language which gave the Table 3 is following:

TestCase1 < - convolutionQuantile(Proposed_Matrix, kernel, x = 0.7)

Table 5. Kernel matrix with diagonal value all 1

1	0	0
0	1	0
0	0	1

Here kernel is the aforesaid matrix, x is numeric vector of probabilities with values in [0,1] which will help the function convolutionQuantile to find the position of quintile 'x' in each cell of the data matrix.

Step 2: **Backpropagation Neural Network with the convoluted data matrix as input neuron**

The description of the experiment using the proposed model of BP ANN is in the following table (Table 6):

Table 6. The experiment description using the proposed steps in tabular form of the BP ANN model

Steps	Description
Model type	As told earlier we are taking a 3-layer backpropagation neural network
Transfer function	$\frac{1}{1+e^{-x}}$
Number of neurons in the hidden layer	10 neurons
Input layer	Here we are taking dataelement3 (E3 as de3), dataelement4 (E4 as de4), dataelement5 (E5 as de5), dataelement9 (E9 as de9) as the input neuron as they have the least range of data variation in their columns
Output layer	One Neuron. This Neuron will state the occurrence of the population of the stem borer

The obtained Neural Network according to the aforesaid proposed method is the following (Fig. 1):

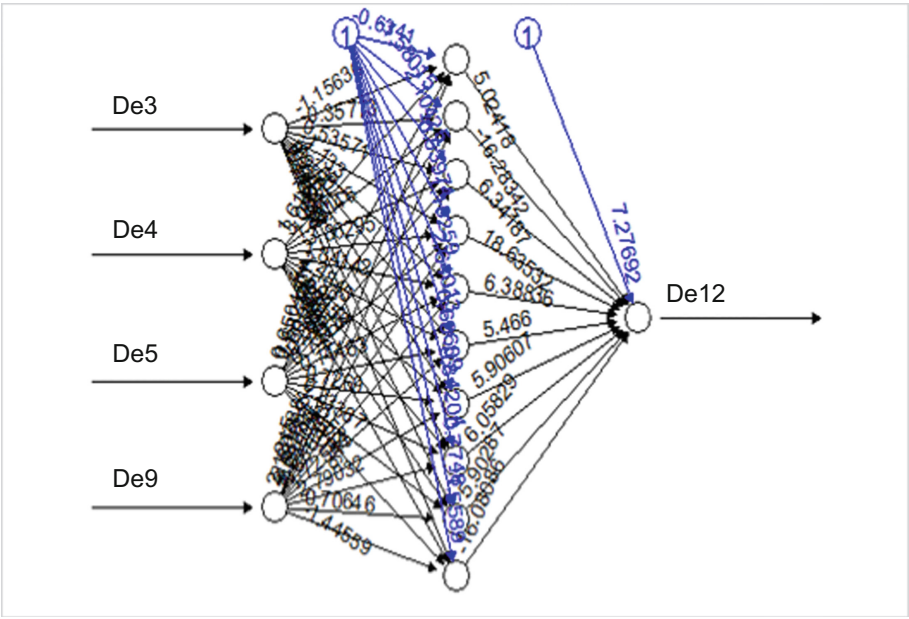

Fig. 1 Obtained Neural Network of the proposed method.

In the proposed method it has been used 50% of the convoluted data as the training set to train the neural network and 50% of the data to test the output of the neural network.

4.2 Comparison with the Existing Model

ANN is an excellent choice to use in prediction. In the time of applying ANN to predict the population occurrence of stem borer, here ANN system tried to imitate the thinking process of humanbeings, such as observing, learning and concluding. Then a model environment is set up to learn and gather information based on appropriate data structures. The existing model depicts these in the following way:

Step 1: Normalize the influencing factors
Step 2: Principal Components Analysis to reduce the dimensions of dataset
Step 3: Backpropagation ANN training with the principal components from the previous steps.

After implementing the existing model the result is in the following table (Table 7):

Table 7. Results of the existing model

Mean square error	0.001148403589
Reached threshold	0.009284247547
Steps	33

In the proposed model the result is in the following table (Table 8):

Table 8. Results of the proposed model

Mean square error	0.000000001675894812
Reached threshold	0.005154583570896686
Steps	229

Therefore, apart from the calculation steps according to the comparison of the error rate the proposed model is better than the existing model. This is shown in the following figure by comparing the intercept of propagated error in each hidden layer in both the existing and proposed model which would eventually lead to calculation of mean square error in predicting the population occurrence. In the following figure it is evident that the proposed model has a lower spread in error in each step compare to the existing model (Fig. 2).

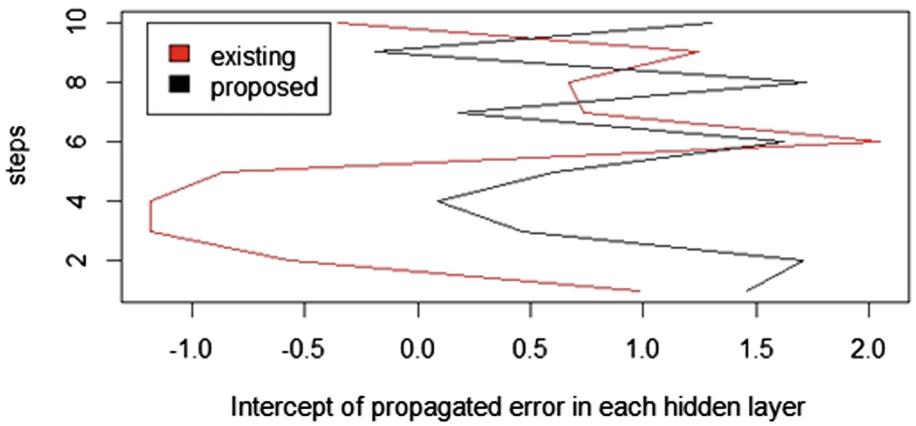

Fig. 2 Comparison of intercept of propagated error in each hidden layer with each interception step in the existing and proposed model

5 Conclusions and Future Work

Anything one can do with a Convolution Neural Network (CNN), one can also do it with a fully connected architecture just as well. But as it has a lot of PPPPI. It is more convenient for use. That means when one has a lot of features (like an image does), using a CNN, one can get comparable learning potential with far fewer parameters. As a result, one can train faster and use less data. Convolutional networks work so well because they exploit an assumption about weight sharing. This is why they only work with data where that assumption holds. As shown in the result analysis portion the size of the database becomes very less compare to the already existed model by using the CNN with compare to Principal Component Analysis. According to Yang et al. [1, 12], on the basis of experience, experiment, and statistical method, there are three main ways to forecast pest occurrences. Anything that is modeled by ANN and has a time series data flow one can also depict it as a CNN. This is why the next approach is to make the prediction model more improved by using the other steps of the CNN along with the BP-ANN network to enhance the performance of the model by utilizing advanced statistical data analytics.

References

1. Yang, L.N., Peng, L., Zhang, L.M., Li-lian, Z., Yang, S.S., et al.: A prediction model for population occurrence of paddy stem borer (Scirpophaga incertulas), based on back propagation artificial neural network and principal components analysis. Comput. Electron. Agric. **68**(2009), 200–206 (2009). https://doi.org/10.1016/j.compag.2009.06.003
2. Rad, M.R.N., Koohkan, S., Fanaei, H.R., Rad, M.R.P.: Application of Artificial Neural Networks to predict the final fruit weight and random forest to select important variables in native population of melon (Cucumis melo. Pahlavan). Sci. Hortic. **181**, 108–112 (2015). https://doi.org/10.1016/j.scienta.2014.10.025

3. Sun, S., et al.: A comparison of models for the short-term prediction of rice stripe virus disease and its association with biological and meteorological factors. Acta Ecol. Sin. **36**, 166–171 (2016). https://doi.org/10.1016/j.chnaes.2016.04.002
4. Günther, F., Fritsch, S.: neuralnet: Training of neural networks. R J. **2**(1), 30–38 (2010)
5. Smith, B.A., Gerrit, H., McClendon, R.W.: Artificial neural networks for automated year-round temperature prediction. Comput. Electron. Agric. **68**(2009), 52–61 (2009). https://doi.org/10.1016/j.compag.2009.04.003
6. Sengar, R.S., Kalpana, S.: Climate Change Effect on Crop Productivity. Taylor & Francis Group LLC, Milton Park (2015)
7. Seyedmohammadi, J., Esmaeelnejad, L., Ramezanpour, H.: Determination of a suitable model for prediction of soil cation exchange capacity. Model. Earth Syst. Environ. **2**, 156 (2016). https://doi.org/10.1007/s40808-016-0217-4
8. Borovykh, A., Bohte, S., Oosterlee, C.W.: Conditional Time Series Forecasting with Convolutional Neural Networks. arXiv:1703.04691v3[stat.ML] 16 October 2017
9. Aaron, C., Cristopher, M., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature **453**, 98–101 (2008)
10. Cai, Z.X., Xu, G.Y.: Artificial Intelligence (AI) and Its Application. Tsinghua University Press, Beijing (1996)
11. Chen, X.J., Wang, C.N., Chen, S.T.: The Principle and Application of Neural Network, pp. 1–3. National Defense and Industry Press, Beijing (1995)
12. Maria, J.D.: Artificial neural networks as an alternative tool in pine bark volume estimation. Comput. Electron. Agric. **48**(3), 235–244 (2005)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**(2012), 1097–1105 (2012)
14. Wang, Z., Yan, W., Oates, T.: Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline. CoRR, abs/1611.06455 (2016)
15. Hagan, M.T., Demuth, H.B., Beale, M.H., De Jesus, O.: Neural Network Design (2 edn, ebook). hagan.okstate.edu/nnd.html