Dynamic Neuro-fuzzy Inference and Statistical Models for Risk Analysis of Pest Insect Establishment

Snjezana Soltic[1, 2], Shaoning Pang[2], Nikola Kasabov[2], Sue Worner[3], and Lora Peackok[3]

[1] Department of Electrical & Electronic Engineering, Manukau Institute of Technology, Manukau City, New Zealand
ssoltic@manukau.ac.nz
[2] Knowledge Engineering & Discovery Research Institute
Auckland University of Technology, Auckland, New Zealand
spang@aut.ac.nz
[3] Center for Advanced Bio-protection Technologies,
Ecology and Entomology Group
Soil, Plant and Ecological Science Division
Lincoln University, Canterbury, New Zealand
Worner@lincoln.ac.nz

Abstract. The paper introduces a statistical model and a DENFIS-based model for estimating the potential establishment of a pest insect. They have a common probability evaluation module, but very different clustering and regression modules. The statistical model uses a typical K-means algorithm for data clustering, and a multivariate linear regression to build the estimation function, while the DENFIS-based model uses an evolving clustering method (ECM) and a dynamic evolving neural-fuzzy inference system (DENFIS) respectively. The predictions from these two models were evaluated on the meteorological data compiled from 454 worldwide locations, and the comparative analysis shows advantages of the DENFIS-based model as used for estimating the potential establishment of a pest insect.

1 Introduction

A variety of methods have been designed to predict the likelihood of pest establishment upon a species introduction into an area [1], [2], [3], [4], [5], [6], [7]. It is observed that, (1) a number of methods have been developed specifically for problems at hand, and therefore have relatively narrow applicability, and (2) usually only one method was applied to a data set, and therefore there is a lack of comparative analysis that show advantages and disadvantages of using different methods on the same data set.

The analysis of the response of a pest to influential environmental variables is often so complex that traditional methods are not very successful. Artificial neural networks have been studied as a promising tool for decision support in ecological research [8], [9]. The studied neural networks are mainly of a multilayer perceptron type that have some drawbacks such as absence of incremental learning, no facility for extracting knowledge (rules) and often, not good generalization [8]. This research

N.R. Pal et al. (Eds.): ICONIP 2004, LNCS 3316, pp. 971–976, 2004.

describes and compares two models for predicting the potential establishment of a pest in new locations using *Planocuccus citri* (Risso), the citrus mealybug, as a case study. The software environment NeuCom (www.kedri.info) was used in the paper for the analysis and the prediction.

2 Experiments

2.1 Data Set

In the experiment, meteorological data compiled from 454 worldwide locations where *Planocuccus citri* (Risso) has been recorded as either present (223 locations) or considered absent (232 locations), were used. Each location is described using a 16-dimensional vector and a class label (present/absent). Note that, the class label for a number of locations from the absent class might be false absent. The pest species may be absent at a location simply because it may never have reached it, and not because the climate is unsuitable for its establishment.

2.2 Problem Definition

The assessment of the establishment potential of any species (response variable) can be formulated by the following: Given a problem space: $D = \{X_1, X_2, \cdots, X_k, Y\}$, where $X_i (i = 1, \cdots, k)$ are data examples from D, and $Y = y_1, y_2, \cdots, y_k$ is the vector under estimation. Suppose $X = x_1, x_2, \cdots, x_l$. The target is to predict Y in terms of X by modeling an estimation function $Y = f(X)$. The estimation function f is then used to make spatial predictions of the response, e.g., to predict the establishment of a pest in a new area following entry.

2.3 Models

Two models are introduced and discussed in this paper: (1) a statistical model, and, (2) a dynamic evolving neural-fuzzy inference system (DENFIS)-based model, which are denoted as Model I and Model II respectively.

These two models have a common probability evaluation module, but very different clustering and regression modules. Model I uses a typical K-means algorithm for data clustering [10], and a multivariate linear regression to build the estimation function. Model II clusters data using an evolving clustering method (ECM) [10] and estimates f by a dynamic evolving neural-fuzzy inference system (DENFIS). The details of the DENFIS can be referenced in [10], [11]. Both models fit response surfaces as a function of predictors in environmental space $E = \{X_1, X_2, \cdots X_k\}$, where $X_i (i = 1, \cdots, k)$ are data examples from D and then use the spatial pattern of predictor surfaces to predict the response in geographical space $G = \{g_1, g_2, \cdots, g_k\}$, where the examples are of type $g_i = (latitude_i, longitude_i)$. Model II is incrementally trainable on new data in contrast to Model I.

We implemented the statistical model to predict the establishment potential as follows.

1. Apply a clustering algorithm to data from the problem space D.
2. Suppose $\{C_1, C_2, \cdots, C_\xi\}$, are clusters from the clustering module. For each cluster $C_i \in \{C_1, C_2, \cdots, C_\xi\}$ calculate the mean vector and establishment potential using:

$$X_i^c = \frac{\sum_{j=1}^{|C_i|} X}{|C_i|}, \quad p_i^c(Y \mid x_1, x_2, \cdots, x_k) = \frac{\sum_{j=1}^{|C_i|} p(y \mid x_1, x_2, \cdots, x_k)}{|C_i|}, i = 1, \cdots \xi. \qquad (1)$$

3. Use $\mathbf{P}^c$ and $\mathbf{X}^c$ to build the estimation function f.
4. Use f to make spatial predictions of the response (e.g., estimate the establishment potential for each location given in the original data set D).

Note that the regression is performed among clusters C, instead of among samples in D. This enables the model to estimate probability without losing the key information among clusters.

The above procedure was repeated using both models. In Model I the K-means module was used for clustering of the original data set D where the number of clusters, iterations and replicates was set to 20, 100 and 5 respectively. In Model II ECM was used for partitioning data D into 20 clusters (the number of clusters can be and was controlled by selecting the maximum distance, *MaxDist*). Thereafter, the multiple linear regression model was used to build the estimation function (Model I):

$$\begin{aligned} y = {} & 0.78017 - 0.52528x_1 - 0.1023x_2 + 4.262e-005x_3 + 0.030326x_4 + 0.0020693x_5 \\ & + 1.0084x_6 - 1.748x_7 + 1.9414x_8 - 0.13537x_9 - 1.1652x_{10} + 0.87642x_{11} \\ & - 0.08011x_{12} - 0.96676x_{13} - 0.078018x_{14} + 1.9266x_{15} - 1.2633x_{16} \end{aligned}$$

In Model II DENFIS was applied to $\mathbf{P}_{ecm}^c$ and $\mathbf{X}_{ecm}^c$. Consequently, we obtained 15 rules, each of them representing the 15 rule nodes created during learning. Those rules cooperatively function as an estimate that can be used to predict the establishment potential of the citrus mealybug at each location.

The first rule extracted is as follows:

Rule 1: if x_1 is f(0.20 0.75) & x_2 is f(0.20 0.70) & x_3 is f(0.20 0.10) & x_4 is f(0.20 0.53) & x_5 is f(0.20 0.33) & x_6 is f(0.20 0.73) & x_7 is f(0.20 0.75) & x_8 is f(0.20 0.76) & x_9 is f(0.20 0.76) & x_{10} is f(0.20 0.72) & x_{11} is f(0.20 0.71) & x_{12} is f(0.20 0.69) & x_{13} is f (0.20 0.69) & x_{14} is f(0.20 0.71) & x_{15} is f(0.20 0.72) & x_{16} is f(0.20 0.71) then

$$\begin{aligned} y = {} & -2.45 - 27.88x_1 - 150.94x_2 - 1.27x_3 - 4.04x_4 + 4.65x_5 - 59.0x_6 + 85.32x_7 \\ & - 19.85x_8 - 29.54x_9 + 72.0x_{10} + 45.41x_{11} - 129.34x_{12} + 203.15x_{13} \\ & + 11.39x_{14} + 12.75x_{15} - 6.59x_{16} \end{aligned}$$

3 Results

In Table 1, we compared the DENFIS-based model with the statistical model on the establishment potential prediction of the citrus mealybug at 24 locations. The first 12 locations were chosen because they were given establishment potential estimates greater than 0.7 by Model I. The second 12 locations were given estimates greater than 0.7 by Model II. Each location is described by a pair of geographic coordinates (latitude, longitude), which is given in column 2. The predictions by the statistical and the DENFIS-based are presented in column 3 and column 4, respectively. For the purpose of the comparison, column 5 records the known establishment status of the pest (presence: 1/absence: 0).

Table 1. Results for 24 selected locations. The correct matches are shown in bold

Location	(Latitude, Longitude)	Model I	Model II	Label
Shaam, Selenge	(50.1, 106.2)	1	**0.45**	0
Saran-Paul', Russia	(64.28, 60.88)	0.87	**0.55**	0
Nape, Laos	(18.3, 105.1)	0.80	**0.42**	0
Bangladesh	(24, 90)	**0.80**	0.65	1
Hacienda Santa Elena	(22.52, -99)	0.75	**0.47**	0
Seoul	(37.6, 127)	**0.74**	0.48	1
Tamanrasset, Algeria	(22.78, 5.52)	0.74	**0.39**	0
Najaf, Iraq	(31.98, 44.32)	0.74	**0.55**	0
Dhubri, India	(26.02, 89.98)	0.73	**0.63**	0
Thailand	(16, 102)	**0.73**	0.55	1
Asuncion, Paraguay	(-25.3, -57.7)	**0.73**	0.60	1
Monclova, Coah.	(26.88, -101.42)	0.72	**0.41**	0
Valencia	(39.5, -0.4)	0.49	**1**	1
Lima	(-12.1, -77)	0.16	**0.87**	1
Torit, Sudan	(4.4, 32.5)	0.42	**0.84**	1
Juba, Sudan	(4.87, 31.6)	0.42	**0.83**	1
Ghana	(8, -1)	0.49	**0.75**	1
Ibadan, Nigeria	(7.4, 3.9)	0.41	**0.75**	1
Rwanda	(-2, 30)	0.41	**0.74**	1
Uganda	(2, 32)	0.47	**0.73**	1
Zhejiang (Chekiang)	(29, 120)	0.27	**0.71**	1
Trinidad	(21.48, -80)	0.29	**0.71**	1
Fujian / Fukien	(26, 118)	0.36	**0.71**	1
Dakar, Senegal	(14.7, -17.5)	0.36	**0.71**	1

Given a threshold value, P_{thr}, for scores greater than P_{thr}, set $P = 1$ representing the pest presence, otherwise set $P = 0$ and the pest is absent. Given a location $g_i = (latitude_i, longitude_i)$ in column 2, if the prediction P_i equals to the true value from the 5th column, then the prediction is matched. As can be seen, Model I gives 4 matches in 24 locations, while Model II gives 20 matches.

In Fig. 1 we carried out another comparison, where establishment potentials of citrus mealybugs from 454 worldwide locations are estimated by the above two prediction models, and their performances were measured by match-degree/threshold-value plots. The match-degree, defined as a ratio between the number of locations with a match and the total number of locations, was assessed over the range $P_{thr} \in [0.4, 0.8]$. As can be seen, although both models have similar accuracy predicting the absence of the pest, Model II slightly outperforms Model I. In the case of the presence of the pest, Model II is better than Model I in that Model II achieves more matches than Model I for each $P_{thr} \in [0.4, 0.8]$. Particularly, when $P_{thr} \geq 0.6$, the two models give a significant difference in accuracy, where Model II accuracy increases to 100% while the accuracy of the Model I drop down to 0%.

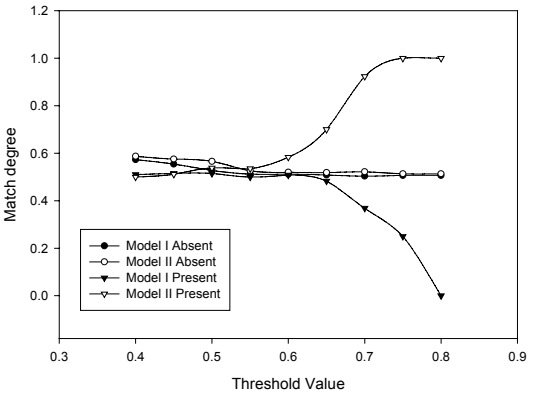

Fig. 1. The accuracy of the models predicting the pest presence or absence at 454 locations expressed in terms of match-degree/threshold-value plots

4 Conclusions

In this paper, we introduced and compared a statistical model and a DENFIS-based model for estimating the potential establishment of pest insects. We used both models in a case study to predict the establishment of the citrus mealybug.

The DENFIS-based model is recommended for on-line prediction applications. If new, yet unseen data becomes available DENFIS will adapt its structure and produce output to accommodate the new input data. During learning, this model creates rules that are useful to researchers who study pest-environmental relationships. The model is preferred because it employs local rather than global clustering, thus the information about pest locations is better conserved in the estimation. This comparative analysis clearly illustrates the advantages of the DENFIS-based model when used for estimating the establishment potential of this particular species of pest insect, and therefore it is a possible new solution for general pest risk assessment.

Acknowledgments

Snjezana Soltic wish to acknowledges the support of this work by the Research Committee of the Department of Electrical and Electronic Engineering at the Manukau Institute of Technology, through the Departmental Research Fund.

References

1. Sutherst R.W., Maywald, G.F. and Bottomley, W.: From CLIMEX to PESKY, a generic expert system for pest risk assessment. *EPPO Bulletin* (1991) **21**:595-608
2. Dentener, P.R., Whiting D.C., Connoly, P.G.: Thrips palmi karny (Thysanoptera: Thripidae): Could it survive in New Zealand? In: Proc. of 55th Conference of New Zealand Plant Protection Society Incorporated (2002) 18-24
3. Dobesberger, E.J.: Multivariate techniques for estimating the risk of plant pest establishment in new environments. Presented at NAPPO International Symposium on Pest Risk Analysis, Puerto Vallarta, Mexico, (2002) Available: http://www.nappo.org/PRA-Symposium/PDF-Final/Dobesberger.pdf , December 2003
4. Dobesberber, E.: Climate based modelling of pest establishment and survival in support of rest risk assessment., In: Annual report 1999-2000, North American Plant Protection Organization (2000) 35-36, Available: http://www.nappo.org/Reports/AnnRep-99-00-e.pdf, December 2003
5. Stynes, B.: Pest risk analysis: methods and approaches. Presented at NAPPO PRA Symposium, Puerto Vallarta, Mexico, (2002) http://www.nappo.org/PRA-Symposium/PDF-Final/Stynes.pdf , December 2003.
6. Baker, R.H.A.: Predicting the Limits to the Potential Distribution of Alien Crop Pests. In: Halman G., Schwalbe, C.P. (eds.): Invasive arthropods and agriculture: problems and solutions. Science Publisher Inc., Enfield, New Hampshire (2002) 208-241
7. Cohen, S.D.: Evaluating The Risk of Importation of Exotic Pests Using Geospatial Analysis and Pest Risk Assessment Model. First International Conference on Geospatial Information in Agriculture and Forestry, Lake Buena Vista, Florida, USA, (1998) http://www.aphis.usda.gov/ppd/evaluating.pdf December 2003
8. Worner, S.P. et. al.: Neurocomputing for decision support in ecological research. Conference on Neurocomputing and Evolving Intelligence, Auckland, New Zealand, 20-21 November 2003 (2003)
9. Gevrey, M., Dimopoulus, I., Lek, S.: Review and comparison of methods to study the contribution of variables in artificial neural network models. In: Ecological Modelling 160 (2003) 249-264
10. Kasabov, N.: Evolving connectionist systems: Methods and applications in bioinformatics, brain study and intelligent machines. Springer-Verlag (2002)
11. Kasabov, N., Song, Q.: Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time-Series Prediction. In: IEEE Trans. on Fuzzy Systems, vol. 10. (2002) 144-154