Artificial Neural Networks to Predict Daylily Hybrids

Ramana M. Gosukonda (Gosukonr@fvsu.edu), Masoud Naghedolfeizi (feizim@fvsu.edu), and Johnny Carter (carterj@fvsu.edu)
Fort Valley State University, Fort Valley, GA 31030

Abstract

Artificial Neural Networks (ANN) were employed to predict daylily (Hemerocalli spp.) hybrids from known characteristics of parents used in hybridization. Features such as height, diameter, foliage, blooming habit, ploidy, blooming sequence were included in the initial training and testing. Data pre-processing was performed to meet the format requirements of ANN. Backpropagation (BP), Kalman filter (KF) learning algorithms were used to develop nonparametric models between the input and output data sets. These networks were compared with traditional multiple linear regression models. Prediction plots for both height and diameter indicated that the regression model had a better accuracy in predicting unseen patterns. However, ANN models were able to more robustly generalize and interpolate unseen patterns within the domain of training.

1. Introduction

Daylilies (*Hemerocalli spp*.) are important in international trade and domestic economy with revenues over millions of dollars per year. Daylily growers are often interested in knowing what pod and pollen parents would result in a desired hybrid, as the seeds bear by the pod parent frequently results in hybrids with different patterns. Further, the growers have to wait at least a year before these hybrids can grow and exhibit patterns. However, with help of traditional and artificial neural network (ANN) models, this information could be provided in advance to growers for efficient daylily breeding. Therefore the objective of this paper is to make a better use of available data for estimating daylily parameters using ANN and/or regression models. Correlation between height and diameter are well observed features in daylily. In this report we discuss the analysis of height data and diameter data.

2. Daylily Data Structure

The data sets used in the study are taken from the daylily database compiled by the American Hemerocallis Society [1]. Though this database has over 56,000 records, because of its rigid format, about 230 sample patterns were manually extracted to spreadsheet format. Out of 230 patterns 175 were used in training and 55 were used in testing. Except scape height and diameter, other parameters such as season, foliage, bloom habit and sequence, and ploidy are in text format. These text data were converted to appropriate numerical values before feeding it into ANN. The text variables were scaled to vary between ±1 for ploidy, habit, foliage, and ±3 for blooming sequence. The data were sorted based on the parent's height to determine the domain of training.

3. Prediction Models

To predict the height and diameter of hybrid daylily, known characteristics of its parents were subjected to two different prediction models; ANN (Figure 1) and linear regression. The multiple linear regression model was selected because visual inspection of the data and correlation factor revealed a nearly linear behavior. The ANN models were developed using a commercially available software system [2]. The inputs to the networks were heights and diameters of both parents with a single output of child's height or diameter. The learning algorithms were Kalman filter (KF) and backpropgation [3, 4]. This network configuration was also suitable for comparing the performance of the two models.

The multiple linear regression models were developed using the following equation:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

where Y_1 is either the height or diameter of the child and X_1, X_2 are the height and diameter of parent one and X_3, X_4 are the height and diameter of parent two.

5. Results and Analysis

Tables 1 and 2 present performance measurement parameters for both ANN and regression models of height and diameter, respectively. The analysis of the results indicates that the regression models provide a better correlation and a smaller standard deviation (std) as compared to BP and KF in the prediction region (testing dataset) (Figures 2 and 3). However, BP and KF performed better in terms of both std and correlation factor in interpolating unseen patterns. The main advantage of neural network models over regression models was the ability to recognize the relationship between input and output data sets without specifying a priori relationship.

Table1. Performance statistics of diameter

Statistics	Linear regression	Back-propagation	Kalman Filter
Correlation*	0.75	0.65	0.69
Std. Dev *	1.41	1.39	1.25
Correlation**	0.77	0.91	0.85
Std. Dev **	1.02	1.09	1.05

* represent prediction data
** represent interpolation

Table2. Performance statistics of height

Statistics	Linear regression	Back-propagation	Kalman Filter
Correlation*	0.55	0.31	0.25
Std. Dev*	5.96	7.70	6.51
Correlation**	0.62	0.90	0.84
Std. Dev**	5.39	1.09	5.97

* represent prediction data
** represent interpolation

Figure 1: A typical diagram of a feedforward neural network model

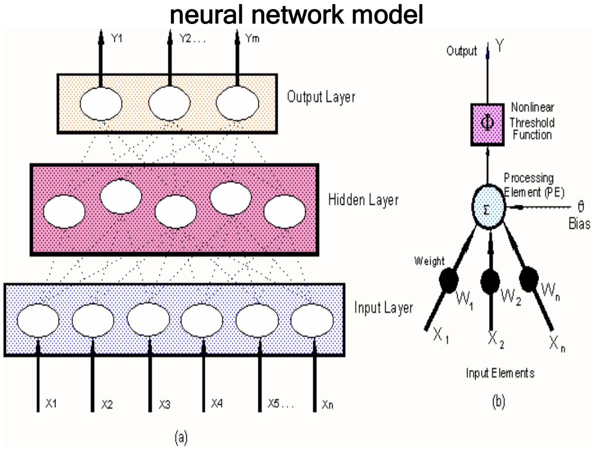


Figure 2: Height plots on prediction data

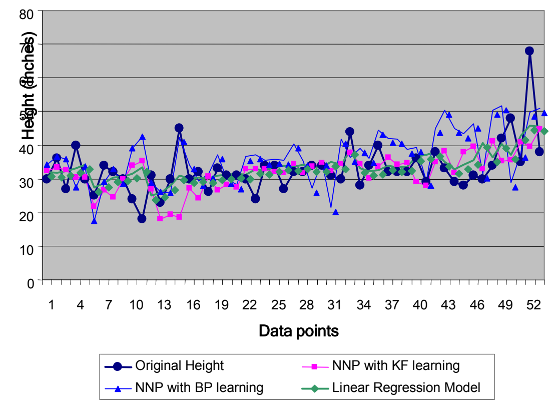
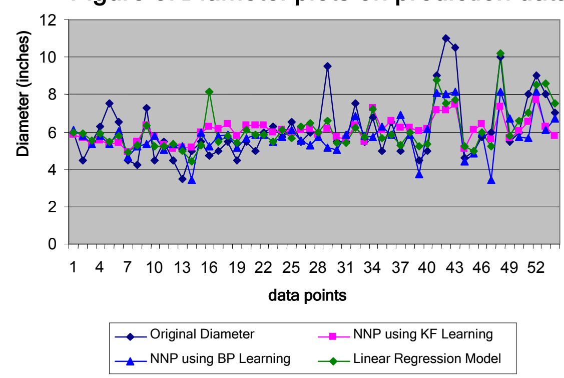


Figure 3. Diameter plots on prediction data



6. Conclusions

This study demonstrated that both neural network and traditional regression techniques could be used to predict features of a daylily hybrid from its parents' features with a relatively good accuracy. However, more patterns are needed to identify which technique would provide a better paradigm for developing prediction models.

7. References

[1] The official electronic checklist of Hemerocallis Cultivar Registrations 1890-2004. 2005. Compiled by American Hemerocallis Society, TN.
[2] Neuralworks Predict version 3.0, Neuralware Company, www.neuralware.com.
[3] Simon Haykin, Neural Networks: A Comprehensive Foundation: 2/e, Prentice Hall, 1999, NJ.
[4]Ruck, D.W., Rogers, S.K., Kabrisky, M., Maybeck, P.S., and Mills, J.P.1992. Comparative analysis of backpropagation and the extended Kalman filter for training multilayer perceptrons. IEEE Transactions on Pattern Analysis and Machine Intelligence. 14(6): 686-691.