Letter to the Editor

A new $R^2$-based metric to shed greater insight on variable importance in artificial neural networks

ARTICLE INFO

Keywords:
- Connection weights
- Garson’s algorithm
- Machine learning

ABSTRACT

Artificial neural networks (ANNs) represent a powerful analytical tool designed for predictive modeling. However, the shortage of straightforward and reliable approaches for calculating variable importance and characterizing predictor–response relationships has likely hindered the broader use of ANNs in ecology. Two such metrics – product-of-connection-weights (PCW) and product-of-standardized-weights (PSW) – have received much attention in the published literature. A recent paper (Fischer, in press, Ecological Modelling) found that PSW was comparable to PCW for retrieving variable importance values in linear models – seemingly overturning the conclusions of Olden et al. (2004, Ecological Modelling) – and that PSW was superior to PCW in nonlinear models. In this paper we call into question the findings of Fischer (in press) and more importantly we explain why neither PCW nor PSW are universally good measures of variable importance. We advance the field by proposing a new permutational $R^2$-based variable importance metric and show that it accurately estimates the proportion of the total variance in the response variable that is uniquely associated with each predictor variable in both linear and non-linear data contexts. By enabling ecologists to measure relative strengths of predictor variables in a transparent and straightforward way this metric has the potential to help widen the use of ANNs in ecology.

Published by Elsevier B.V.

1. Introduction

Artificial neural networks (ANNs) have witnessed greater application in ecology in recent decades because they are perceived to overcome many of the difficulties commonly associated with ecological data (Olden et al., 2008). Included are the advantages of modeling non-linear associations, not requiring specific assumptions concerning the distributional characteristics of the independent variables (i.e., nonparametric), and accommodating variable interactions without a priori specification. ANNs perform complex pattern recognition and prediction tasks by adaptively learning and mapping input variables to output variables via a series of interconnected nodes akin to neurons in the brain (Rumelhart et al., 1986; Bishop, 1995; Ripley, 1996). However, adding model complexity to enhance predictive power often comes at the cost of reduced explanatory insight into the ecological relationships being explored (Loé-Décarie et al., 2014).

With the aim of illuminating the “black box” perception of ANNs a variety of methods have been developed to quantify predictor (input) variable importance in neural networks (reviewed in Olden and Jackson, 2002; Gevrey et al., 2003; Olden et al., 2004). In a study published in Ecological Modelling, Olden et al. (2004) provided a comprehensive evaluation of the accuracy of different approaches in retrieving true variable importance in a simulated dataset where predictor variables were linearly associated with the response variable. This study found that the product-of-connection-weights method (hereafter, PCW) was the best performing metric while Garson’s (1991) product-of-standardized-weights (hereafter, PSW) method performed the worst.

In a recent paper, also published in Ecological Modelling, Fischer (in press) reported that PSW was a better measure of variable importance in ANNs compared to PCW when applied to datasets exhibiting nonlinear relationships between predictor and response variables. Interestingly, in the simulation analysis by Fischer (in press), PSW correctly retrieved variable importance ranks in all simulated datasets (4 functional forms: 1 linear; 2 quadratic; 1 interaction-term; for each functional form, 100 datasets with sample size $n = 500$ were evaluated) whereas PCW was reported to be 100% accurate in the linear datasets but only 16–52% accurate in the other 3 nonlinear datasets.

Although we are encouraged that discussion of how explanatory insight is gleaned from ANNs continues in the literature, regrettably Fischer’s (in press) analysis was problematic in a number of ways. First, it was simply impossible to replicate the simulation experiments because Fischer (in press) failed to provide information regarding the response variable’s error structure. Second, it is unclear how variable importance values of PSW and PCW can be related to beta-weights of the correct linear regression model (see Fig. 1 in Fischer, in press); neither Garson’s (1991) original formulation of PSW nor Olden et al.’s (2004) proposal of PCW asserted this relationship. Therefore, one cannot interpret the relative importance values of both PSW and PCW presented by Fischer (in press); or put another way, there was no theoretical benchmark upon which to make a comparison. Third, even if we

http://dx.doi.org/10.1016/j.ecolmodel.2015.06.034
0304-3800/Published by Elsevier B.V.
assume beta-weights are a legitimate benchmark for PSW and PCW importance values, the relative importance of two variables \( x_1 \) and \( x_2 \) does not equal the ratio of beta-weights of \( x_1^2 \) and \( x_2 \) which was used by Fischer (in press) to evaluate the two quadratic datasets; the latter ratio measures the relative importance of \( x_1^2 \) and \( x_2 \) (not \( x_1 \) and \( x_2 \)). Fourth, a 100% accuracy rate is exceedingly rare in simulation experiments owing to stochastic noise in the simulated dataset. Because the error structure is not known, we were not able to explain this result in the context of the simulation design. Moreover, a 100% accuracy in determining variable importance ranks in dataset 4 (functionual form: \( y = x_1 + x_2 + 0.1 \times x_3 \), where \( x_1 \) and \( x_2 \) have the same theoretical ranks) required the implausible result of identical variable importance values for the two variables in all simulations.

The problems listed above and the high uncertainty around the variable importance values associated with PCW prompted us to investigate the legitimacy of both connection weight-based metrics, PCW and PSW. First, we performed a simulation study in which we assessed the utility of PCW and PSW in retrieving variable importance ranks from simulated datasets that follow the linear model structure used by Fischer (in press). However, unlike Fischer (in press), we performed a more robust investigation using a variety of sample sizes and overall effect sizes to more completely evaluate the generality of his findings (i.e., 100% rank accuracy). We did not repeat Fischer’s experiment for nonlinear and interaction-term models because the beta-weight ratio used in his simulation experiment does not correctly measure variable importance. Second, we performed a more detailed dissection of a one hidden-layer ANN to explain why neither PCW nor PSW are universally good measures of variable importance rank or value. Third, in recognition of the weaknesses of both PCW and PSW, we propose a simple and generalizable \( R^2 \)-based variable importance measure and evaluate its performance against an appropriate benchmark.

2. Methods

2.1. Simulation experiment

We simulated datasets from the linear function used by Fischer (in press) (Eq. (1)):

\[
y_i = 1 \times x_{i1} + 0.5 \times x_{i2} + 0.1 \times x_{i3} + e_i
\]

(1)

where \( y_i \) is ith observation of the response (output) variable, \( x_{i1}, x_{i2}, x_{i3} \) are ith observations of the three predictor variables, and \( e_i \) is the stochastic error term of the ith observation that follows a Gaussian distribution with mean = 0 and standard deviation \( \sigma \).

We simulated 9 sample size–effect size combinations (3 sample sizes \( \times 3 \) effect sizes), each of which comprised 1000 datasets. In comparison, Fischer (in press) simulated a single sample size (\( n = 500 \)) for a single effect size (unknown) examining just 100 datasets. In our experiment, we varied the overall effect size (i.e., strength of association between the predictor variables and response variable or overall \( R^2 \)) by changing the standard deviation of the error term (\( \sigma = 0.3, 1, 2 \), corresponding to overall model \( R^2 \approx 0.93, 0.56, 0.24 \), respectively). Sample sizes in each dataset were \( n = 100, 500, \) and 1000. For each dataset, we first generated \( x_1, x_2, \) and \( x_3 \) from a uniform distribution bounded between 0 and 1. We then centered and scaled each predictor variable so that mean = 0 and variance = 1 before simulating the response variable following Eq. (1).

We analyzed the 1000 datasets in each sample size–effect size combination separately using ANNs fitted by the backpropagation algorithm. We used the same network architecture as Fischer (in press): 3 input nodes, 3 hidden nodes in a single hidden layer, and 1 output node. The activation function in the output node was linear (as appropriate for the data) whereas that in the hidden nodes was sigmoidal (logistic), which is standard for ANNs. As the decay parameter was unknown in Fischer (in press), we chose to perform 10-fold cross validation to search for the optimal parameter value (from 5 candidate values 0.0001–1) that minimized root-mean-square-error (RMSE) for each simulated dataset. The model with the optimal value was used to determine variable importance. Specifically, we wanted to evaluate whether we could replicate the 100% PSW rank accuracy in Fischer (in press).

2.2. Functional decomposition of a simple ANN

Using a simple ANN with 2 input nodes, 2 hidden nodes, and 1 output node as an example, we explain how predictor (input) variables and neural connection weights combine to produce predictions of the response (output) variable. By comparing the functional relationship between the input variables and the output variable with the PSW and PCW methods, we evaluated the legitimacy of these two weight-based methods in determining relative variable importance.

2.3. A new \( R^2 \)-based measure of relative variable importance

After revealing methodological weaknesses of PSW and PCW, we propose a new measure of relative variable importance for ANNs based on the reduction of the variance explained by the model (\( R^2 \)) by permuting each predictor variable in turn. Permuting a given predictor variable breaks the association between the predictor variable and the response (if present) and decreases the overall \( R^2 \) of the model. The magnitude of the reduction in \( R^2 \) when a given predictor variable is permuted reflects the strength of association between that predictor variable and the response. The rationale behind this metric follows the permutation accuracy importance metric used in random forest models (Strobl et al., 2008).

The permutational relative variable importance (pRV1, \( i \)) of the \( i \)th predictor variable \( x_i \) is (Eq. (2)):

\[
pRV1_{xi} = R^2_{\text{obs}} - R^2_{\text{perm,xi}}
\]

(2)

where \( R^2_{\text{obs}} \) is the \( R^2 \) of the ANN model fitted to the observed predictor and response variables, \( R^2_{\text{perm,xi}} \) is the \( R^2 \) of the ANN model fitted to a modified dataset where \( x_i \) is permuted, and \( R^2_{\text{perm,xi}} \) is the mean (or alternatively, median) value of \( R^2_{\text{perm,xi}} \) after \( m \) permuted datasets. \( R^2 \) was calculated as 1 – residual sum-of-squares/total sum-of-squares. This framework can be generalized to incorporate different measures of model fit in place of \( R^2 \) such as RMSE for regression and classification accuracy for classification tasks. Because pRV1 is based on the permutation of predictor variables, it can be applied to datasets comprising both continuous and categorical predictor variables.

We tested the utility of pRV1 by conducting a simulation experiment that fitted ANNs to simulated datasets based on the three functional forms used by Fischer’s (in press) (Eqs. (3)–(5)):

Linear:

\[
y_i = 1 \times x_{i1} + 0.5 \times x_{i2} + 0.1 \times x_{i3} + e_i
\]

(3)

Quadratic:

\[
y_i = -1 \times x_{i1}^2 + 0.5 \times x_{i2} + 0.1 \times x_{i3} + e_i
\]

(4)

Interaction-term:

\[
y_i = 1 \times x_{i1} \times x_{i2} + 0.1 \times x_{i3} + e_i
\]

(5)

where \( y_i \) is ith observation of the response (output) variable, \( x_{i1}, x_{i2}, x_{i3} \) are ith observations of the three predictor variables, and \( e_i \) is the stochastic error term of the ith observation that follows a Gaussian distribution with mean 0 and standard deviation \( \sigma \).

This experiment had the same design as the one presented in Section 2.1. We simulated 9 sample size–effect size combinations, each comprising 500 datasets for each function. The effect sizes, dataset sample sizes, and variable generation procedure follows Section 2.1.
For each simulated dataset, we determined optimal values for the decay parameter (5 values: 0.0001–1) and the number of hidden nodes (3 values: 2, 3, 4) by performing 10-fold cross validation. To guard against false convergence on local minima that is characteristic of machine learning approaches (Olden et al., 2008; preliminary analyses indicated that false convergence affected quadratic and interaction-term datasets), we added an algorithm to accept an ANN model only when the range of $R^2$ obtained from four consecutive models were within a value of 0.01. With the optimal ANN model structure, we calculated $pRVI$ and $pRVI$ rankings for each dataset and compared it to a known benchmark: relative importance as measured by the reduction in $R^2$ when fitting the correct model [via least-squares (LS) fitting] after permuting a given predictor variable versus fitting the correct model to the original dataset. For analyses involving the interaction-term datasets, a correct variable importance ranking was defined by $pRVI$ of $x_1$ and $x_2 > pRVI$ of $x_3$.

2.4. Software and code

All analyses were conducted in R 3.2.0 (R Core Team, 2015). Following Fischer (in press), ANNs were fitted in the nnet package (Ripley & Venables, 2015). PSW and PCW were calculated with functions provided in the NeuralNetTools package (Beck, 2015). We used the doParallel package (Revolution Analytics, 2014) to parallelize simulations. Due diligence necessitates that all code is made freely available when reporting on methodological comparisons using simulated data. For this reason we provide example R code for simulations, convergence algorithm, and $pRVI$ in Supporting Code A1.

3. Results

3.1. Simulation experiment

Contrary to Fischer (in press), neither PSW nor PCW produced 100% accurate variable importance ranks when applied to datasets with an underlying linear structure (Table 1). Notably, PCW was more accurate than PSW in all 9 sample size–effect size combinations we examined. The accuracy of PSW and PCW generally declined with decreasing sample size and increasing stochastic noise (i.e., error $\sigma$) as expected. However, the decline in PSW accuracy was much more pronounced than that of PCW. When $\sigma = 2$ ($R^2 \sim 0.26$) and $n = 100$, PSW gave correct variable rankings for only
34% of the datasets whereas PCW ranked 76% of these datasets correctly. The suboptimality of PCW and PSW in ranking variables cannot be attributed to the effect of stochastic noise on linear regression beta-weights. In all but the lowest sample size—effect size combination, rankings derived from beta-weights were ≥ 99.7% accurate. In other words, although almost all simulated datasets (even with stochastic noise) reflected the correct variable importance structure, PCW and PSW were unable to retrieve it accurately.

The variation in PCW and PSW values was much higher than the benchmark beta-weight values in all sample size—effect size combinations (Fig. 1 and Fig. A1).

### 3.2. Functional decomposition of a simple ANN

We used a simple ANN with 2 input nodes, 2 hidden nodes, and 1 output node to illustrate the relationship between input variables to output variables via neural weights (Fig. 2). The output values from the 2 input nodes ($IN_1$ and $IN_2$) are directly connected to the input (predictor) variables $x_1$ and $x_2$ via an identity link; therefore $IN_1 = 1 \times x_1$ and $IN_2 = 1 \times x_2$. These values are fed forward to the 2 hidden nodes; the net input (or activation) for each node is the sum of its input values (i.e., input values are $IN_1$, $IN_2$, and $BIAS_H = 1$) multiplied by their associated connection weights ($w$ and $b$; see Fig. 1). The activation values ($A_1$ and $A_2$) for hidden nodes 1 and 2 are therefore (Eqs. (6) and (7)):

$$A_{H_1} = w_{IN_1,H_1} \times IN_1 + w_{IN_2,H_1} \times IN_2 + 1 \times b_{H_1}$$

(6)

$$A_{H_2} = w_{IN_1,H_2} \times IN_1 + w_{IN_2,H_2} \times IN_2 + 1 \times b_{H_2}$$

(7)

The output values ($H_1$ and $H_2$) from the hidden nodes are obtained by applying a transfer function to $A_{H_1}$ and $A_{H_2}$. The transfer function in the hidden nodes is often sigmoidal; logistic or hyperbolic tangent functions are commonly used. We use a logistic transfer function in our example (Eqs. (8) and (9)):

$$H_1 = \frac{1}{1 + \exp(-A_{H_1})}$$

(8)

$$H_2 = \frac{1}{1 + \exp(-A_{H_2})}$$

(9)

The transfer function in the hidden layer is never linear because a network with such a transfer function would be reduced to a single layer network (without hidden layers) that can only deal with linearly separable problems (i.e., like linear regression) (Bishop, 1995).

Next, the output values from the hidden nodes are fed forward to produce a net activation value ($A_{OUT}$) into the output node which is in turn put through a transfer function to produce the overall output value ($OUT$; corresponding to response variable $y$) (Eqs. (10) and (11)):

$$A_{OUT} = w_{H_1,OUT} \times H_1 + w_{H_2,OUT} \times H_2 + 1 \times b_{OUT}$$

(10)

$$OUT = 1 \times A_{OUT}$$

(11)

We use an identity (linear) output transfer function here but a sigmoidal transfer function could also be used depending on the nature of the response variable in the dataset.

Combining Eqs. (6)–(11), the relationship between the input variables ($x_i$) and the response variable ($y$) can be explicitly written down as (Eq. (12)):

$$y = \frac{w_{H_1,OUT} \times H_1}{1 + \exp\left[-\left(w_{IN_1,H_1} \times x_1 + w_{IN_2,H_1} \times x_2 + b_{H_1}\right)\right]} + \frac{w_{H_2,OUT} \times H_2}{1 + \exp\left[-\left(w_{IN_1,H_2} \times x_1 + w_{IN_2,H_2} \times x_2 + b_{H_2}\right)\right]} + b_{OUT}$$

(12)

As seen from Eq. (12), the relative effect of the predictor variables $x_i$ on the output $y$ cannot be calculated by comparing products of the connection weights across the network—the method used in both PSW (Garson, 1991) and PCW (Olden et al., 2004) approaches. This is because the marginal change in the output of a given hidden node ($H$) resulting from one unit change in $x_i$ depends on the bias...
term $b_H$ even when the connection weights $w$ are the same. Consequently, the connection weights do not measure effect sizes in a consistent manner.

3.3. A new $R^2$-based measure of relative variable importance

In general, our new $R^2$-based metric $p_{RVI}$ reflected the actual reduction in $R^2$ when the correct least-squares (LS) model is fitted to permuted datasets (Figs. 3–5). The variable importance rankings retrieved by $p_{RVI}$ were $\geq$95% accurate in all but 2 functional model–sample size–effect size combinations (Table 2).

Although generally high, the accuracy of $p_{RVI}$ varied depending on the functional model underlying datasets, sample size and effect size categories. The position of the points relative to the $1:1 (p_{RVI}:actual R^2$ reduction) line indicated that $p_{RVI}$ was slightly more accurate in linear and interaction–term datasets (Figs. 3 and 5) than datasets with a quadratic term (Fig. 4). $p_{RVI}$ accuracy increased as sample size increased (Figs. 3–5, bottom to top), but interestingly, decreased with increasing effect size (Figs. 3–5; right to left). We did not present figures from analyses involving intermediate sample ($n = 500$) and effect ($\sigma = 1$) sizes because they followed the trends described above.

When overall effect size was extremely high ($\sigma = 0.3$ corresponding to mean $R^2 > 0.90$), $p_{RVI}$ tended to underestimate the importance of the most highly correlated variable(s) and overestimate the importance of subordinate variables(s). This bias was much reduced when the overall effect size was low ($\sigma = 2$ corresponding to mean $R^2 < 0.30$).

4. Discussion

Despite performing extensive simulations comprising a wide range of sample and effect sizes, we could not replicate Fischer’s (in press) findings that PSW (Garson, 1991) was 100% accurate in ranking variable importance in linear datasets. Our simulations confirmed Olden et al.’s (2004) findings that PCW outperformed PSW in such datasets; however, neither of these two approaches were consistently accurate.

Our theoretical analysis demonstrated that both PSW and PCW are not entirely faithful approaches to determine variable importance in ANNs, thus explaining the poor empirical performance of these connection weight-based metrics. The reason is that the effect exerted by 1 unit of connection weight is unequal across hidden nodes because the activation function is sigmoidal. When the bias weight is a large positive or negative value, a 1 unit increase in connection weight increases the output value by a smaller value than when the bias weight is close to zero.

If there is no theoretical basis for both PSW and PCW, why were they effective in determining variable importance values and ranks in previous studies (e.g., Olden et al., 2004; Fischer, in press)? In Olden et al.’s (2004), the simulation datasets on which PCW was evaluated had a very high overall effect size. This might have caused bias weights to be similar across hidden nodes resulting in comparable connection weights. The products of comparable connection weights were therefore able to approximate importance rankings (but probably not values). We could not speculate on the reasons for Fischer’s (in press) findings owing to a lack of information on the error structure of the data generating model as well as various
Fig. 4. Permutational relative variable importance (pRVI) values versus actual $R^2$ reduction from fitting the correct quadratic least-squares regression model after permuting each given variable in datasets with (a) $n = 1000$, $\sigma = 0.3$, (b) $n = 1000$, $\sigma = 2$, (c) $n = 100$, $\sigma = 0.3$, (d) $n = 100$, $\sigma = 2$. Points that fall on the 1:1 (black) line are pRVI values that accurately estimate actual $R^2$ reduction.

Fig. 5. Permutational relative variable importance (pRVI) values versus actual $R^2$ reduction from fitting the correct interaction-term least-squares regression model after permuting each given variable in datasets with (a) $n = 1000$, $\sigma = 0.3$, (b) $n = 1000$, $\sigma = 2$, (c) $n = 100$, $\sigma = 0.3$, (d) $n = 100$, $\sigma = 2$. Points that fall on the 1:1 (black) line are pRVI values that accurately estimate actual $R^2$ reduction.
other aspects involving the ANN parameterization. However, it is possible that Fischer’s (in press) simulated datasets exhibited very large effect size ($R^2 > 1$).

Our $R^2$-based variable importance metric ($pRVI$) is a possible alternative to connection weight-based approaches. By measuring the proportion of variance in the response that is uniquely explained by each variable, it is analogous to the squared semipartial correlation in linear regression. Importantly, $pRVI$ accurately retrieved variable importance values and rankings as defined by the unique proportion of variance explained in linear as well as nonlinear datasets. One pitfall of this method is that the relative importance of the strongest variables is slightly underestimated and importance of the weakest variables is overestimated when sample sizes are low and effect sizes are extremely high. However, this bias is likely to be negligible for most ecological analyses because it is extremely rare for ecological datasets to exhibit such high overall effect sizes (model $R^2 > 0.90$).

In conclusion, connection weight-based approaches to estimating variable importance are problematic. They lack theoretical support and do not perform well when applied to datasets with small–moderate effect and sample size. We propose a permutation-based $R^2$-based metric ($pRVI$) which is general for different types of input variables and is accurate in measuring variable importance in linear as well as non-linear datasets. The use of $pRVI$ together with association visualization tools (see Lek et al., 1996; Olden and Jackson, 2002) can help ecologists gain better explanatory insight into their study systems using ANNs.

**Acknowledgment**

Financial support was provided by a H. Mason Keeler Endowment to University of Washington. We thank one anonymous reviewer for commenting on the manuscript.

**Appendix A. Supplementary data**

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2015.06.034

**References**


Xingli Giam,*

Julian D. Olden**

* Corresponding author.

** Corresponding author. Tel.: +1 206 616 3112.

E-mail addresses: xgliam@uw.edu (X. Giam),
olden@uw.edu (J.D. Olden).

25 May 2015