CONSORTIUM

History:
Received: December 13, 2016
Accepted: December 23, 2016
First Published: December 28, 2016
Collection year: 2016
Confirmation of publication: Published

Identifiers and Pagination:
Year: 2016
Volume: 1
First Page: 18
Last Page: 21
Publisher ID: 19257430.6.18
DOI: 10.21065/19257430.6.18

Corresponding author:
Ecevit EYDURAN PhD
Igdir University, Agricultural Faculty, Department of Animal Science, Biometry Genetics Unit, Igdir-Turkey.
E.: ecevit.eyduran@gmail.com

Citation:
Ecevit E. The possibility of using data mining algorithms in prediction of live body weights of small ruminants. Adv Cal Anal 2016, Vol. 1, *p* 18-21

Editorial

THE POSSIBILITY OF USING DATA MINING ALGORITHMS IN PREDICTION OF LIVE BODY WEIGHTS OF SMALL RUMINANTS

Ecevit EYDURAN

Igdir University, Agricultural Faculty, Department of Animal Science, Biometry Genetics Unit, Igdir-Turkey

Keywords: Data Mining Algorithm, Body Weight Prediction, CART, MARS, CHAID, ANN

The main purpose of the sheep production is to improve profitability of yield traits such as meat, milk and wool obtained per animal. In this respect, selection is a remarkable tool for achieving genetic improvement and attaining better qualified offspring as to the quantitative traits. In obtaining of superior offspring according to a quantitative trait like live weight, the conservation of indigenous genetic sources and the detection of the breed standards, animal breeders take into account indirect selection criteria with the help of high genetic correlation coefficients between live weight and morphological traits. Moreover, the prediction of live body weight from some zoometrical (morphological) characteristics measured simply in farm animals is an important subject for developing prosperous animal breeding systems and in practice, regulating management conditions [1; 2]. A simple way to find out appropriate feed amount, medicinal dose and price of an animal farm is to predict live body weight from effective morphological traits. The predictive accuracy depends on choosing powerful statistical approaches. Among those, there is multiple linear regression, which leads analysts to make biased parameter estimates with multicollinearity problem occurring as an outcome of very strong Pearson correlation coefficients between morphological traits as predictors of body weight [3]. A good alternative is, in general, to use Ridge Regression Analysis instead. However, Ridge regression can produce unreliable outcomes [4]. More effective alternatives to remove multicollinearity problem are available, such as using scores of factor analysis and principal component analysis for multiple regression analysis technique [5; 6]. Predictors are exposed to factor or principal component analysis as one of multivariate analysis techniques and new uncorrelated predictors are used to predict the body weight without multicollinearity problem [6].

Recent studies show that the most effective alternatives in the body weight prediction are data mining algorithms. Among these algorithms, CART (Classification and Regression Tree), CHAID (Chi-Square Automatic Interaction Detector) and Exhaustive CHAID construct a regression tree structure that can be interpreted easily by researchers. CART tree-based algorithm recursively products binary splits by partitioning a subset into two small subsets until achieving the strongest Pearson coefficient in body weight trait between observed and predicted values. CHAID algorithm recursively uses multi-way splitting in regression tree construction for the strongest Pearson coefficient as a model quality criterion [7]. In the CHAID algorithms, there are three stages, merging, splitting and stopping and the Bonferroni adjusment is available in the estimation of adjusted P values. The last two stages are the same; however, Exhaustive CHAID algorithm employs an

Funding:
The authors received no direct funding for this research.

Competing Interests:
The authors declare no competing interests

Additional information is available at the end of the article.

exhaustive procedure in order to merge any similar pairs until obtaining merely a single pair in regression tree structure. CHAID algorithms implement F significance test when a response variable (body weight) is continuous. In this situation, the tree diagram constructed for a continuous response variable in CART and both CHAID algorithms is called the regression tree, otherwise named as the classification tree. CHAID algorithms become automatically active to prune the redundant structures in the regression tree diagram. However, in the CART algorithm, analysts should activate a pruning option.

Usability of Artificial Neural Networks (ANNs) algorithms as more sophisticated approaches in the prediction of body weight is scarce [7]. To reveal the complicated relationship between a response variable (body weight) and other input variables (predictors), ANNs, functioning like human brain and consisting of input, hidden and output layers, are the best choice. However, it is extremely difficult to interpret their outputs compared with the tree-based data mining algorithms. In this respect, ANNs are also called as black boxes.

For researchers who aim to predict an equation for body weight, application of MARS (Multivariate Adaptive Regression Splines) data mining algorithm which is unavailable in literature should be preferred. More importantly, MARS, a non-parametric regression statistical technique to get linear piecewise functions and to evaluate high order interactions between predictors, is used to reveal more complex relationships between sets of more-than-one dependent variables and predictors with the aid of pruning option. Compared to other statistical approaches mentioned above, MARS provides a much higher predictive performance in prediction problems. For this reason, MARS can be applied to RSM data consisting of more-than-one dependent variables and predictors in agricultural and medical sciences. This type of application is absent in literature.

Several model evaluation criteria are recommended in testing and comparing predictive performances of the statistical approaches addressed above [7].

a) Pearson correlation coefficient (r) between the actual and predicted BW values,

b) Root-mean-square error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{ip}\right)^2}$$

c) Mean error (ME) given by the following equation:

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{ip}\right)$$

d) Mean absolute deviation (MAD):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_{ip}\right|$$

e) Standard deviation ratio (SD_{ratio}):

$$SD_{ratio} = \frac{s_m}{s_d}$$

f) Global relative approximation error (RAE):

$$RAE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_{ip}\right)^2}{\sum_{i=1}^{n} y_i^2}}$$

g) Mean absolute percentage error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_{ip}}{y_i}\right| \cdot 100$$

h) Coefficient of Determination

$$R^2 = \left[1 - \frac{\sum_{i=1}^{n}\left(y_i - y_{ip}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2}\right]$$

i) Adjusted Coefficient of Determination

$$R^2_{ADJUSTED} = \left[1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}\right]$$

Where:

n is the number of animals in a set, *k* is the number of model parameters, y_i is the observed value of a response variable (Body weight), y_{ip} is the predicted value of the response variable (Body weight), s_m is the standard deviation of the model residuals, s_d is the standard deviation of the response variable (Body weight) and $\overline{Y}$ is the mean of the response variable (Body weight).

The best model should have the greatest Pearson coefficient, R^2 and adjusted R^2 and the lowest RMSE, MAD, MAPE and RAE. SD ratio should become equal to the value less than 0.40 for a good fit in model, and for very good fit, the ratio should be equal to the value less than 0.10 [7].

Consequently, researchers generally prefer more understandable and interpretable statistical approaches. In the scope of regression analysis, the most fundamental purpose is to minimize residuals expressed as differences in body weight between observed and predicted values or to maximize Pearson correlation coefficient between observed and predicted values, obtained by statistical analysis approach, in the body weight.

References

1. Eyduran, E., Zaborski, D., Waheed, A., Celik, S., Karadas, K., Grzesiak, W., 2017. Comparison of the Predictive Capabilities of Several Data Mining Algorithms and Multiple Linear Regression in the Prediction of Body Weight by Means of Body Measurements in the Indigenous Beetal Goat of Pakistan. Pakistan J. Zoology in press.
2. Karadas, K., Tariq, M., Tariq, M.M, Eyduran, E., 2017. Measuring Predictive Performance of Data Mining and Artificial Neural Network Algorithms for Predicting Lactation Milk Yield in Indigenous Akkaraman Sheep. Pakistan J. Zool., 49(1):1-7.
3. Tariq, M.M., Rafeeq, M., Bajwa, M.A., Awan, M.A., Abbas, F., Waheed, A., Bukhari, F.A., Akhtar, P., 2012. Prediction of body weight from body measurements using regression tree (RT) method for indigenous sheep breeds in Balochistan, Pakistan. The J. Anim. Plant Sci. 22(1):20-24.
4. Jahan, M., Tariq, M.M., Kakar, M.A., Eyduran, E., and Waheed, A., 2013. Predicting Body Weight From Body And Testicular Characteristics of Balochi Male Sheep in Pakistan Using Different Statistical Analyses. The Journal of Animal & Plant Sciences, 23(1): 14-19.
5. Eyduran, E., Karakus, K., Karakus, S., Cengiz, F., 2009. Usage of factor scores for determining relationships among body weight and body measurements. Bulgarian J. Agric. Sci., (15): 374–378.
6. Khan, M.A., Tariq, M.M., Eyduran, E., Tatliyer, A., Rafeeq, M., Abbas, F., Rashid, N., Awan, M.A., Javed, K., 2014. Estimating body weight from several body measurements in Harnai sheep without multicollinearity problem. J. Anim. Pl. Sci., (24): 120-126.
7. Ali, M., Eyduran, E., Tariq, M.M., Tirink, C., Abbas, F., Bajwa, M.A., Baloch, M.H., Nizamani, A.H., Waheed, A., Awan, M.A., Shah, S.H., Ahmad, Z., Jan, S., 2015. Comparison of artificial neural network and decision tree algorithms used for predicting live weight at post weaning period from some biometrical characteristics in Harnai Sheep. Pakistan J. Zool., vol. 47(6):1579-1585.