Weight Analysis and Prediction of Yak

Xiao-feng QIN¹, Chen ZHANG¹, Zi-jie SUN¹, Yu-an ZHANG¹,*, Ren-de SONG² and Mei-yun DU¹

¹Department of Computer Technology and Applications, Qinghai University, 810016, Xining, China
²Yushu Prefecture Animal Husbandry and Veterinary Station, 815000, Yushu, Qinghai, China

*Corresponding author

Keywords: Yak weight, Weight prediction, K-means clustering, Pearson correlation, Linear regression prediction, Support vector machine prediction.

Abstract. This paper aims to predict the weight of yak by analyzing the relationship between body weight and body size characteristics of yak in Three-Rivers Source area of Qinghai Province, and to guide the growth, development and breeding of yak in this area. In this paper, we selected 303 healthy yaks in Yushu Prefecture of Qinghai Province, the weight and body size (body height, body length, chest circumference, tube circumference) data of yak at different age groups were measured by facilities of weighbridge, measuring stick and tape measure. The data set is divided into training set and test set by random function, cluster the data use k-means clustering algorithm to observe the data distribution. At the same time, Pearson correlation analysis is used to calculate the correlation between body weight and body size, and the training set is analyzed by multiple linear regression and support vector machine prediction. The results show that the average relative error between the predicted value and the true value of the support vector machine is 5.72%, and the average relative error of the multiple linear regression and the true value is 6.52%, which means the support vector machine prediction algorithm can predict the weight of the yak better.

Introduction

The Three-Rivers Source area is located in the southern plateau of Qinghai Province with a total area of 363,000 square kilometers, which unique geographical conditions and environmental factors have nurtured a wealth of rare animals. It is the region with the highest concentration of ecological diversity in high altitude areas around the world, the alpine meadow is the main grassland type of the hinterland and main part of the Qinghai-Tibet Plateau. As one of the main livestock species in the Three-Rivers Source region, yak is not only a symbol of herders' production, life, social economy and national culture, but also constitute the main body of the alpine meadow ecosystem with Tibet sheep[1].

Yak is a special kind of livestock that can adapt to the cold climate and can make full use of alpine grassland resources to produce meat, milk, cow hair and etc., which can survive in the harsh environment where other livestock are difficult to adapt [2]. Due to the ups and downs of the alpine pastures, traffic is extremely inconvenient, it is very difficult to measure the weight of the yaks under natural conditions. Most of the regions because the backward economic and poor infrastructure so that can only estimate the body weight [3].The body size index of livestock reflects the size and body structure of the livestock, development and other conditions, which is closely related to body weight, and with the convenient measurement and small error to establish the body size index to estimate the weight of the yak for the development of animal husbandry in the region has important significance. It is an important means to study the growth and development, breeding, production performance and feeding standards of yak, and is the main indicators to measure the selection of breeding [4].

This study analyzes data by k-means clustering, Pearson correlation analysis, multiple linear regression, and support vector machine methods in data mining. Programming with python to analyze the collected data comprehensively. The model was evaluated and the results of the two methods were compared and analyzed, which laid the foundation for the subsequent related research. And Figure 1 shows the picture of the yak.
Materials and Methods

Data Collection

In this study, 303 healthy yak in Yushu Prefecture of Qinghai Province was selected as the research object. The researchers from Yushu Animal Husbandry and Veterinary Station collected the body weight and body size data of the yak using the weighbridge, measuring stick and tape measure that are measured as shown in Figure 2.

The height of the yak (TG) is the vertical distance between the highest point of the armor and ground contact point that measured by measuring stick; body oblique length (TXC) indicates the distance from the shoulder to the end of the point of the buttocks; the chest circumference (XW) represents the circumference of the vertical axis of the posterior scapula of the shoulder blade measured by the tape measure; the circumference of the tube (GW) indicates the circumference of the thinnest part of the tube bone, which usually from the lower left anterior leg to the upper one third. The weight of the yak value is the measured by the weighbridge after the yak has been fasted in 12 hours.

Data Processing

The collected data is divided into 0.5, 1.5, 2.5, 3.5 and 4.5 years old, write a random function to divide the data as training set and test set in 4:1 ratio by python.

Algorithm Implementation

Multiple Linear Regression Algorithm

Multiple linear regression is an important method in multivariate statistical analysis and is widely used in social, economic, technological and many natural science research, which studies how to effectively organize and analyze randomly affected data and make inferences and predictions on the issues examined [5].

In the actual problem, a certain variable $Y$ is related to other variables $x_1, x_2, ..., x_{p-1}$, but because of this correlation and mechanism is not clear. Therefore, it can only be said that the value of $Y$ is determined by the part of value of $x_1, x_2, ..., x_{p-1}$. On this occasion, considered the value of $Y$ consists as two parts, a part that can be determined by $x_1, x_2, ..., x_{p-1}$, which is a function of independent variables, the other part is the effect of many unconsidered factors, which is considered as a random error[6], denoted as $\mu$. Thus, the linear regression equation is:

$$ Y = f(x_1, x_2, ..., x_{p-1}) + \mu $$  \hspace{1cm} (1)

Regression analysis uses the observation data of $Y$ and $x_1, x_2, ..., x_{p-1}$, and determines the function $(x_1, x_2, ..., x_{p-1})$ under certain assumptions of the error term, when $f(x_1, x_2, ..., x_{p-1})$ is a linear function of $x_1, x_2, ..., x_{p-1}$, the general formula of multiple linear regression model:

$$ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \mu $$  \hspace{1cm} (2)
Where $\beta_0, \beta_1, ..., \beta_{p+1}$ are unknown constants, called regression coefficients, which are solved by least squares method in analysis, $Y$ is called a dependent variable or a response variable, $x_1, x_2, ..., x_{p-1}$ are called independent variables or explanatory variables. $\mu$ is called the random error term, and $E(\mu) = 0$ is assumed in this model.

Multivariate linear regression analysis was performed on yaks of all ages from 0.5 to 4.5 years old, which is shown in Table 1.

<table>
<thead>
<tr>
<th>Age</th>
<th>Constant</th>
<th>TG</th>
<th>TXC</th>
<th>XW</th>
<th>GW</th>
<th>RMSE(10)</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>-94.25</td>
<td>0.77</td>
<td>-0.33</td>
<td>0.72</td>
<td>-0.36</td>
<td>4.28</td>
<td>0.841</td>
</tr>
<tr>
<td>1.5</td>
<td>-178.64</td>
<td>0.26</td>
<td>0.81</td>
<td>1.35</td>
<td>1.62</td>
<td>9.24</td>
<td>0.859</td>
</tr>
<tr>
<td>2.5</td>
<td>-231.76</td>
<td>0.85</td>
<td>0.82</td>
<td>1.23</td>
<td>2.35</td>
<td>10.67</td>
<td>0.866</td>
</tr>
<tr>
<td>3.5</td>
<td>-357.96</td>
<td>1.46</td>
<td>1.29</td>
<td>1.16</td>
<td>4.64</td>
<td>12.35</td>
<td>0.880</td>
</tr>
<tr>
<td>4.5</td>
<td>-447.95</td>
<td>1.96</td>
<td>1.28</td>
<td>1.52</td>
<td>3.94</td>
<td>13.66</td>
<td>0.871</td>
</tr>
</tbody>
</table>

The first column in Table 1 indicates the calves of different ages. Constant indicates the constant term $\beta_0$, TG, TXC, XW, and GW in the multiple linear regression represent the regression parameters corresponding to the explanatory variables. RMSE (10) indicates the error of testing the model using 10-fold cross-validation. $R^2$ is the ratio of the squares sum of the regression and the squares sum of the total deviations, which indicating the ratio of the squares sum of total deviations that can be explained by the squares sum of the regressions. The value of $R^2$ is between 0~1. The closer to 1, the better the regression effect. It is generally considered that the fit of model when it is higher than 0.8. It can be seen from that the $R^2$ value of each age groups are more than 0.8, which indicating that the multiple linear regression model works well [7-8].

The following regression model can be obtained by bringing the parameters in the table into the general formula of the regression model.

$$
0.5 : \quad Y = -94.25 + 0.77 \times TG + 0.33 \times TXC + 0.72 \times XW + 0.36 \times GW \\
1.5 : \quad Y = -178.00 + 0.26 \times TG + 0.81 \times TXC + 1.35 \times XW + 1.62 \times GW \\
2.5 : \quad Y = -231.76 + 0.85 \times TG + 0.82 \times TXC + 1.23 \times XW + 2.35 \times GW \\
3.5 : \quad Y = -357.96 + 1.46 \times TG + 1.29 \times TXC + 1.16 \times XW + 4.64 \times GW \\
4.5 : \quad Y = -447.95 + 1.96 \times TG + 1.28 \times TXC + 1.52 \times XW + 3.94 \times GW
$$

### Support Vector Machine Algorithm

Support vector machine is a machine learning method based on statistics, theoretical VC dimension theory and structural risk minimization principle. It shows many unique advantages in solving small samples, nonlinear and high pattern recognition problems, and to a great extent overcomes the problems of “dimensionality disaster” and “over-learning” [9]. In addition, it has a solid theoretical foundation and a straightforward mathematical model. Therefore, it has been widely developed in the fields of pattern recognition, regression analysis, function estimation, and time series prediction.

Find a regression function $f(x)$ on $\mathbb{R}^n$ for a given sample like $\{(x_1, y_1), (x_2, y_2), ..., (x_i, y_i)\}$ (where $x \in \mathbb{R}$, $x \in \mathbb{R}$, $i=1.2.3....$), there is always a predicted output value $Y_i$ corresponding to any input $X_i$, and the standard deviation corresponding to the output value of each input sample and the predicted value of the output is required to not exceeding $\varepsilon$. The regression function should be as smooth as possible. And the above problem can be described by equation (1) [10].

$$
\min_{w,b,e}(w,e) = \frac{1}{2} \omega^T \omega + \frac{1}{2} \gamma \sum_{k=1}^{N} e_k^2 \quad \text{s. t.} \quad y_k = \omega^T \varphi(x_k) + b + e_k, k = 1, ..., N
$$

Here $y_k$ is not a label indicating the category in the support vector machine, but $y$ is the prediction function in $y = f(x)$, and the same, using the Lagrange multiplier method:
\[ L(\omega, b, e; a) = J(\omega, e) - \sum_{k=1}^{N} \alpha_k \{\omega^T \varphi(x_k) - b + e_k - y_k\} \quad (5) \]

Further, derivation for \( \omega, b, e, a_k \) respectively, the following derivations are available. By simplification, it can be changed to solve the linear equations of equation.

\[
\begin{bmatrix}
0 & 1^T_v \\
1_v & \delta + I/y
\end{bmatrix}
\begin{bmatrix}
b \\
a
\end{bmatrix}
= \begin{bmatrix}
0 \\
y
\end{bmatrix}
\quad (6)
\]

Where \( \delta \) is a kernel function, which is to map a nonlinear problem to a high-dimensional feature space, making it a linear problem. The matrix form can be expressed as:

\[
\delta_{kl} = \varphi(x_k)^T \varphi(x_l) = K(x_k, x_l), \quad k, l = 1, \ldots, N
\quad (7)
\]

Therefore, a regression prediction function can be obtained:

\[
y(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b
\quad (8)
\]

Result and Discussion

Using the multiple linear regression and support vector machine algorithm to estimate the yak weight, the relative error is calculated and the average value is compared to the prediction ability of the algorithm. Based on the analysis of section 3.2 and 3.3, the training set selected by the random function is regarded as the training set, and the body weight of the yak is regarded as the dependent variable, all the body size features that have strong correlation with body weight are selected as independent variables. The prediction results are as shown in Fig.3 and Fig.4.

Fig.3 is a comparison of the predicted values of the multiple linear regression algorithm with the true values. The results are measured by the average of the relative errors. It can be found that the predicted value has a high degree of fitting and the prediction result is comparatively ideal, and the average value of the relative error is 6.52%. Fig.4 shows the comparison between the predicted value of the support vector machine algorithm and the true value, the average value of the relative error is 5.72% and more ideal.

![Figure 3. Comparison of multiple linear regression predicted values and real values.](image-url)
Conclusion

This paper takes the weight and body size data of 303 healthy yak in Yushu Prefecture of Qinghai Province as the research object, and fills in the missing data by professionals for data pre-processing. It can be found that the weight of yak at different ages has a high correlation with body size through Pearson correlation analysis, which is suitable for regression analysis. Therefore, the data were analyzed by multiple linear regression algorithm. The body weight is used as the response variable, body height, body length, chest circumference and tube circumference are explanatory variables, and regression parameters are obtained separately for modeling. The multivariate linear regression was verified by the method of ten-fold cross-validation, and the error was within the acceptable range. Support vector machines have advantages for small-scale data sets. The support vector machine algorithm is used to analyze the data and finds that it supports better performance in predicting yak weight.

Acknowledgment

This work is supported by Science and Technology Project in Qinghai Province (No. 2017-ZJ-717); Yushu Yak High-efficiency Breeding Technology Integration and Demonstration Project (Beijing counterpart assistance).

References


