Research on Modern Higher Education 1, 01002 (2017)
DOI: 10.24104/rmhe/2017.01.01002

An analysis on epidemiological characteristics of brucellosis based on the regression analysis

Jing Qiu
Huazhong Agricultural University, Wuhan, Hubei, China

ABSTRACT: Epidemic situation prediction and prevention concern national defense and people's livelihood. How to control and prevent typical infectious diseases globally is an important subject for scientific researchers. As a zoonosis, brucellosis as well as its prevalence and transmission range cannot be made light of. Epidemiological characteristics of the disease are modeled precisely in this paper. Through the integration of infection sources and infection routes, the author expands the dimension of infection sources and gives comprehensive description and explanation on the probability of human and animal illness. The strategy of deep learning is applied in the regression analysis based on the classical theory of statistics. Accuracy and robustness of the model are improved through the self-learning of the deep learning algorithm on the regression model and regression parameters. This paper provides a reliable theoretical support and planning guide for epidemic prevention and control policies.

Keywords: brucellosis; regression analysis; deep learning; precision modeling

1 INTRODUCTION

Brucellosis is a zoonosis caused by body invasion of bacteria of the genus Brucella. The epidemic situation of brucellosis rebounds worldwide in recent years. The prevention and prediction of epidemic situation of infectious diseases has always been a hot research subject, which plays a vital role in national defense and people's livelihood. The prevention and control of typical infectious diseases is the top priority in recent years. In Increasing Human Brucellosis and Risk Factors Contributing to Its Spatial and Temporal Distribution in China, Yinjun Li carries out a feature modeling of the national morbidity of brucellosis and analyzes epidemiologic features of human brucellosis in China through the regression method and the analytical method of descriptive epidemiology. He also analyzes influencing factors of the regional difference of spatial distribution of human brucellosis in China through the Poisson regression model and carries out an epidemic situation drill through numerical simulation. Methods adopted by Yinjun Li provide valuable information for the epidemic situation risk evaluation of human brucellosis in future [1]. In his An Analysis on the Regional management mode of animal epidemic disease in China and its application, Zhongli Wang not only takes into account problems and influencing factors of animal epidemic diseases control by combining the development status of animal husbandry and the situation of disease prevention and control but also analyzes existing problems of animal epidemics regionalization. He proposes that regionalized management is an effective approach of preventing and controlling animal diseases in China. It provides a guiding direction for China's infectious diseases prevention and treatment in terms of policy [2]. To sum up, most studies on brucellosis involve analyses based on the classical statistics theory and a great deal of work has gained results. However, the transmission speed and the range of infectious diseases have changed a lot with population mobility and economic development. It is an important subject for scientific researchers that how to establish a model with high robustness and accuracy that is suitable for the modern society. In this paper, the author uses the regression analysis algorithm of the classical statistics method optimized through the intelligent algorithm to analyze quantitatively the cross-species transmission of regionally distributed brucellosis. Self-learning is carried out on model parameters through deep learning so as to obtain a model of disease prevention and control with high robustness and accuracy.

2 MODEL AND METHOD

2.1 *Overview on the regression analysis*

In this paper, the author uses the regression analysis algorithm of the classical statistics method to analyze quantitatively the cross-species transmission of regionally distributed brucellosis. Influencing factors of the outbreak of brucellosis and the prevention and control of this epidemic disease can be obtained through the establishment of the related model of human and animal. With the advent of the era of big data, data increases explosively in terms of dimension and volume. The substantial development of information science technology is of great importance for promoting the progress of epidemiology. The application of intelligent algorithm in biological field contributes a lot to accurate analyses of researches in specific fields in the era of big data. Based on the classical algorithm, the author introduces deep learning system to learn regression parameters. The fitting model of brucellosis epidemiological characteristics with high robustness and accuracy is obtained through deep learning.

The author collects information of brucellosis issued by the government and data from epidemic prevention stations and papers as the data support, estimates parameters through the deep learning algorithm, and explains and verifies problems through the established model. Classical models in the regression analysis contain one-dimensional regression models and multi-dimensional regression models. Self-learning on model parameters is carried out in this paper through deep learning, which optimizes model parameters to obtain an ideal model with multiple correlation coefficient, standard deviation, F value and P value as fitness functions.

Commonly used models of nonlinear regression are:

$$\log(Y) = \alpha + \beta \log(X) + \varepsilon$$

$$\log(Y) = \alpha + \beta X + \varepsilon$$

$$\log(\frac{Y}{1-Y}) = \alpha + \beta X + \varepsilon$$

Y is the variable value of regression; X is the independent variable; ε is a random error value; α, β are regression factors.

As for multi-dimensional problems, the suitable multiple regression model is:

$$Y = f(x_1, x_2, x_3, \cdots x_n) + \varepsilon$$

For a simple model, a linear equation is sufficient to depict its essence. The general formula of the linear model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

2.2 *The principle of deep learning*

Deep learning is normally constructed as the model of $m \times n \times l$. Vector quantities are typed in layer m; matrixes are calculated in layer n; vector quantities are printed in layer l. Dimensions of the parameter matrix are set that θ_1 is the first array and θ_2 is the second array. Parameters $\{\theta_1, \theta_2, \ldots \theta_n\}$ are link coefficients between typed vector quantities and the intermediate calculated matrix as well as link coefficients between the intermediate calculated matrix and printed vector quantities [3].

After the construction of the network structure, the network should be calculated forward and cost function should be printed and output. Through the cancellation of bias option, the cost function can be simplified as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}[-y_k^{(i)} \log((h_\theta(x^{(i)}))_k) - (1-y_k^{(i)}) \log(1-(h_\theta(x^{(i)}))_k)]$$

It can be seen that $h_\theta(x^{(i)})$ can be calculated through connections between various links. Here, K=6 is the label value of the global mapping; $h_\theta(x^{(i)})_k = a_k^{(3)}$ is the kth trigger function among printed vector quantities. For the convenience of calculation, the original label is mapped and judged according to the value of {0,1}.

$$y = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad y = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad y = \cdots, \quad y = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

In the original matrix, if the mapping value of $x^{(i)}$ is label5, the corresponding $y^{(i)}$ (calculated through cost function) is the vector of the dimension of the classification number. Here, $y_5 = 1$ and other elements are $y_i = 0$.

Each instance in the training set is mapped and output through the consequent mapping and then collected together [4].

Normally it is necessary to add bias in network calculation. The cost function can be transformed as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}[-y_k^{(i)} \log((h_\theta(x^{(i)}))_k) - (1-y_k^{(i)}) \log(1-(h_\theta(x^{(i)}))_k)]$$
$$+\frac{\lambda}{m}[\sum_{j=1}^{25}\sum_{k=1}^{400}(\theta_{j,k}^{(1)})^2 + \sum_{j=1}^{10}\sum_{k=1}^{25}(\theta_{j,k}^{(2)})^2]$$

Inverse calculation is carried out after the consequent mapping. Inverse calculation is the core of network learning. The gradient of the network function can be obtained through the inverse calculation. Through the gradient calculation, the network can be minimized and adopts algorithms like {fmincg, GA, AP}. Bias can be eliminated by seeking the partial derivative of parameters. Correct derivative calcula-

tion values can be verified through the calculation gradient. Then, bias options are added for training.

The mapping function is set as sigmoid(), the specific expression of which is:

$$sigmoid(z) = g(z) = \frac{1}{1+e^{-z}}$$

Calculate the partial derivative and the solution form of computer is:

$$g'(z) = \frac{d}{dz}g(z) = g(z)(1-g(z))$$

It is able to obtain $J(\theta)$ and cost function through the forward calculation of the network. The inverse calculation is carried out subsequently. Influencing factors are randomly initialized through the inverse calculation so as to break the symmetry inside the matrix. The inverse calculation thought is: as for a group of data ($x^{(t)}$, $y^{(t)}$) in the training matrix, the stimulation of the whole network and $h_\theta(x^{(i)})$ can be obtained by calculating the forward transmission value. After the calculation, calculate $j \subset l$ layer by layer and point by point to obtain the difference value $\delta_j^{(l)}$, which stands for the explanatory component of the whole misjudgment.

As for the output values of the last array of vector quantity, $\delta_j^{(l)}$ can be calculated. As for the intermediate link nodes, $\delta_j^{(l)}$ of the layer can be obtained through the weighted calculation of the last layer, namely:

$$L+1 \mapsto L$$

$$R^{L+1} \Rightarrow R^L$$

Features hidden in the data can be mined through the construction of a deep learning model. After the construction of the network[3], the model can be solved through MATLAB program. Data can be divided into a training set, a verifying set and a testing set. Input data into the model and map independent variables of the training matrix, namely $x^{(t)} \mapsto a^{(1)}$. Carry out the forward calculation so as to obtain values of the next group {$z^{(2)}$, $a^{(2)}$, $z^{(3)}$, $a^{(3)}$}. The stability can be ensured by adding bias.

Set the output item as:

$$\delta_k^{(3)} = (a_k^{(3)} - y_k)$$

Here, $y_k \subset \{0,1\}$

Set the intermediate calculation layer as:

$$\delta^{(2)} = (\theta^{(2)})^T \delta^{(3)} .\times g'(z^{(2)})$$

Obtain $\Delta^{(l)}$ through the accumulation gradient and eliminate $\delta_0^{(2)}$ during the calculation.

$$\Delta^{(l)} = \Delta^l + \delta^{(l+1)}(a^{(l)})^T$$

In network calculation, the extreme value searching of $J(\theta)$ should be verified mathematically. Expand $\theta^{(1)}$ and $\theta^{(2)}$ to obtain a joint vector quantity for the convenience of calculation [5].

Calculate the extreme value of $J(\theta)$ through the hypothesis function $f_i(\theta)$, namely:

$$f \mapsto \frac{\partial}{\partial \theta_i} J(\theta)$$

It is needed to verify that:

$$\theta^{(i+)} = \theta + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \xi \\ \vdots \\ 0 \end{bmatrix}, \quad \theta^{(i-)} = \theta - \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \xi \\ \vdots \\ 0 \end{bmatrix}$$

Whether the partial derivative solution is correct or not can be determined by judging the following formula works or not:

$$f_i(\theta) \approx \frac{J(\theta^{(i+)}) - J(\theta^{(i-)})}{2\xi}$$

Introduce bias if there is no error:

$$\begin{cases} \frac{\partial}{\partial \theta_{i,j}^{(l)}} J(\theta) = D_{i,j}^{(l)} = \frac{1}{m}\Delta_{i,j}^{(l)} + \frac{\lambda}{m}\theta_{i,j}^{(l)} \cdots forj = 1 \\ \frac{\partial}{\partial \theta_{i,j}^{(l)}} J(\theta) = D_{i,j}^{(l)} = \frac{1}{m}\Delta_{i,j}^{(l)} \cdots forj = 0 \end{cases}$$

Here:

$$\theta^{(l)} = \begin{bmatrix} \theta_{1,0}^{(i)} & \theta_{1,1}^{(i)} & \cdots \\ \theta_{2,0}^{(i)} & \theta_{2,1}^{(i)} & \\ \vdots & & \ddots \end{bmatrix}$$

Gradient calculation and verification are carried out after the calculation of the price mapping matrix of the network. The algorithm completes the above steps and then starts the learning of network parameters. A favorable parameter set can be obtained through the learning of fmincg function. In this paper, the author uses the fmincg function of MATLAB itself to carry out parameter optimization [6].

3 ANALYSIS PROCESS AND CONCLUSION

It is believed in classical pathology that the infection source of this infectious disease is livestock. Most research subjects focus on one or only a few creatures. These subjects are one-sided and have limitations. Through the integration of infection sources and infection routes, the author expands the dimension of infection sources, increases infection routes and gives comprehensive description and explanation on the

probability of human and animal illness. Discrete variables are adopted in this paper. Morbidity of sheep, cows and pigs, meat-packing, infection from eating and human-animal contact infection are taken as independent variables, which are respectively represented by $X1$, $X2$, $X3$, $X4$ and $X5$. Human morbidity is the dependent variable. Self-learning on model parameters is carried out in this paper through deep learning, which optimizes model parameters with multiple correlation coefficient, standard deviation, F value and P value as fitness functions [7].

The three kinds of livestock are divided in line with regional identities. The eigenvalue is the combination of the regional frequency of disease development and the radiant scope of influence. The visualization of regional identities is presented in Figure 1.

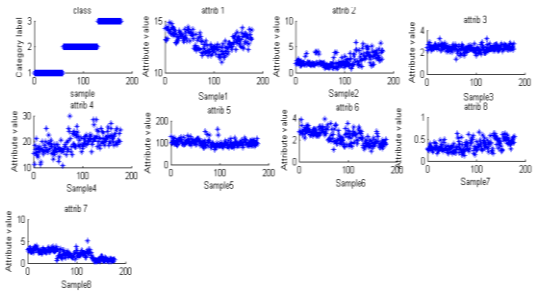

Figure 1. The visualization of regional identities

The three kinds of livestock have different frequency of disease development and influence scope in various regions. Thus, it is suggested to select multiple regions for data analyses instead of the argumentation of a single region or a single factor. Otherwise, it is not comprehensive and brings negative impacts on epidemic situation analyses and prevention.

In this paper, the network structure is a 6-12-3 model with variable discrete values as the input and independent variables of the fitness function as the output. The convergence value is acquired after 95 times of training. The convergence curve is presented in Figure 2.

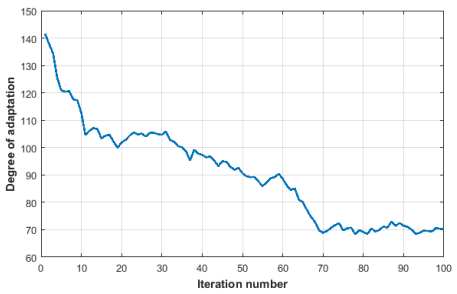

Figure 2. Convergence curve

The parameter selection process can be obtained through visualization. The 3D image of the parameter selection process is presented in Figure 3 and 4.

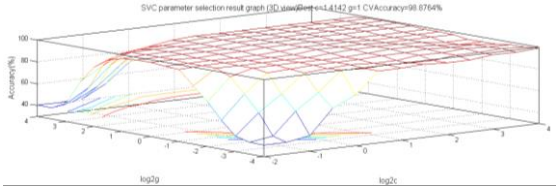

Figure 3. 3D image of parameter selection

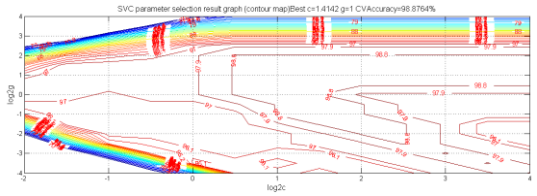

Figure 4. Chorisogram of parameter training

It is able to finally obtain practical models and model parameters through deep learning's self- selection of regression model and self-learning between parameters. In this paper, one-dimensional models and multi-dimensional models are trained together. As for one-dimensional models, the optimal fitness function can be obtained from the selection of five commonly used classical models. As for multi-dimensional models, a maximum dimension is determined for limitation. In this paper, the maximum dimension is 3. Numbers of model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$ (a general form for solution), $\log(Y) = \alpha + \beta\log(X) + \varepsilon$, $\log(\frac{Y}{1-Y}) = \alpha + \beta X + \varepsilon$ and $\log(Y) = \alpha + \beta X + \varepsilon$ are respectively original model 1, original model 2, original model 3, original model 4 and original model 5. Rating diagrams are presented in Figure 5.

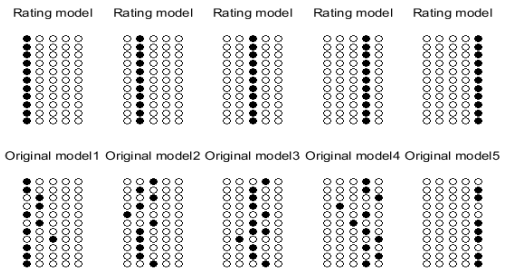

Figure 5. Rating diagrams of models

Through learning and screening of deep learning on the regression model, the optimal fitness function, as well as the sub-optimal fitness function, is a multiple regression model, of which parameters are: $\beta_1 = 0.0027X_1 + 0.123X_2 + 0.115X_3$. The sub-optimal function is a one-dimensional model, of which the model and parameters are: $\log(Y) = \alpha + \beta\log(X) + \varepsilon$. Here, α

and β are respectively -6.1231 and 0.52. DW test values of the optimal model are shown in Figure 6.

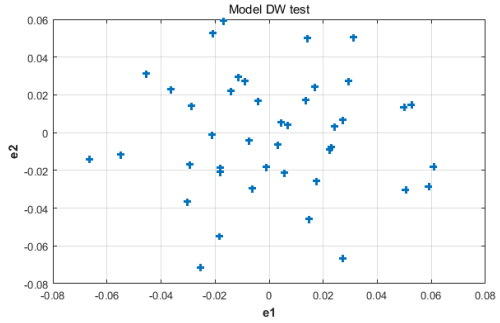

Figure 6. DW test values

It can be seen from the figure that the scatter diagram of errors is distributed in four quadrants evenly. There is no correlation in random errors. The figure of raw residuals and confidence interval and the figure of residuals with eliminated abnormal points and confidence interval are presented in Figure 7 and 8.

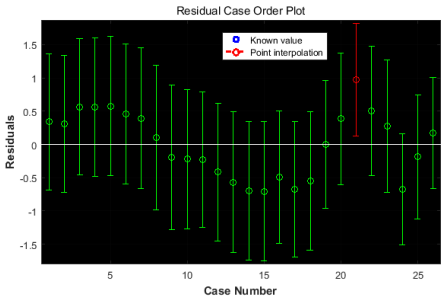

Figure 7. The figure of raw residuals and confidence interval

Figure 8. The figure of residuals with eliminated abnormal points and confidence interval

It can be seen from the empirical analysis on the model obtained through deep learning that the model has a good fault-tolerant ability for data. It is also concluded that the probability of human infection is positively correlated with the three animals. The correlation between human and sheep is the largest. The time-varying interrelation between human infection morbidity and the three animals can be acquired through the training of data of different years. Thus, morbidity and epidemic situation can be controlled perfectly.

4 DISCUSSION AND EXPECTATIO

In this paper, the author successfully introduces the intelligent algorithm into a traditional analysis field by applying deep learning algorithm in the analysis model of epidemiological characteristics of brucellosis. An ideal model is obtained through an empirical analysis. New algorithms can be also introduced to improve the model. However, the stability of these algorithms is still questionable. With the advent of the era of big data, it is a challenging and important subject that how to describe and apply mass data accurately. The field of epidemic control and epidemic surveillance can be optimized and promoted through data acquisition, data processing and data mining.

REFERENCES

[1] Xiao, G., Zhang, G,Q., Sun,Y.X., Zhu, J.L. & Ma, Z. J. 2013. Epidemic dynamics on semi-directed complex networks, *Mathematical Biosciences*, (2).

[2] Hou, Q., Sun, X.D., Zhang, J., Liu, Y.J., Wang, Y.M. & Jin, Z. 2013. Modeling the transmission dynamics of sheep brucellosis in Inner Mongolia Autonomous Region, China, *Mathematical Biosciences*, (1).

[3] Zhang, X., Song, R., Sun, G.Q. & Jin, Z. 2014. Global dynamics of infectious disease with arbitrary distributed infectious period on complex networks. *Discrete Dynamics in Nature and Society*.

[4] Xie, P.T., Eric, P.X. 2013. Multi-Modal Distance Metric Learning. *International Joint Conference on Artificial Intelligence.*

[5] A. Frome, G. Corrado, J. Shlens et al. 2013. Devise: A deep visual-semantic embedding model. NIPS.

[6] Kiros R, Zemel R, Salakhutdinov R. 2014. Multimodal neural language models. *Journal of Machine Learning Research*.

[7] Bengio Y, Courville A, Vincent P. 2013. Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.