Artificial Neural Network Analysis of the Impact of Sample Output Accuracy

Yu Zhong Zhang[1,a], Shao Yun Song[1]

[1]*School of information Technology and Engineering,Yuxi Normal University, Yuxi,Yunnan,China*
[2] *School of information Technology and Engineering,Yuxi Normal University, Yuxi,Yunnan,China*

Abstract. The impact of artificial neural network model output precision technology widespread attention. Quality sample study of neural network output accuracy is not much affected, most of the research is the structure (number of layers and the number of nodes), the impact of this paper to analyze samples of artificial neural network output for the neural network, to improve the output of neural network accuracy is important.

1 Introduction

The neural network to prediction areas, there are many problems to be solved. One of the most prominent problem is there is no standard method to determine the most suitable one neural network structure. Since the parameters that affect neural network prediction capabilities are many. This paper most commonly used BP neural network, quality of the sample neural network prediction accuracy is analyzed in detail and research, and on this basis, to give specific examples of the samples after optimization.

2 Effects of sample quality on the network

Neural networks samples are divided into training sample and testing sample, the quality of the training samples to some extent determines the prediction accuracy. First, there is a big difference between the mean of training samples and the samples to be predicted, the prediction error will be increased as the long-term training. Secondly, the training error increases with training samples will be predicted and the difference between the sample mean increases. Again, with increasing training error would be predicted between the training sample and the sample variance difference increases [1]. Here is the literature [1] method of sample quality analysis.

The neural network prediction error in artificial [1] is $e = e_m + e_t + e_r$, where e is the prediction error, e_m is model error, It was established by the regression model and the differences caused by the actual system, e_t is the final training error, e_r is random errors in the artificial neural network training and prediction process . The existence of e_m and e_t are inevitable, and e_r is $e_m = e_f + e_d$, where e_f is the error between the actual output and predicted output values, which reflects the sample quality; e_d by embedding an incorrect dimension error caused it can be the appropriate number of input neurons by selecting eliminated.

In order to evaluate the quality of training samples, according to e_f proposed indicators "consistent degree". defines a pseudo-distance DCT_{P-D} in paper [1], but pseudo-distance calculation is quite complex and difficult to achieve for analysis and applications of sample quality. The following analysis of the impact on the prediction accuracy of the training sample covariance ratio statistic.

Assume training samples is θ, $\hat{\theta}$ is the output of neural networks, $\hat{\theta}_i$ is the neural network output excluded from eection θ of i-th data points . The ratio statistic covariance of the *i*-th excluded data points is $CR_i = \left|\hat{\text{cov}}(\hat{\theta})\right| / \left|\hat{\text{cov}}(\hat{\theta}_i)\right|$, It shows the effect on the accuracy of the neural network output removed after the i-th data point, From the accuracy portrayed the importance of the *i*-th data point. The value of $\left|CR_i - 1\right|$ greater then the impact of θ_i on the prediction accuracy (neural network output) greater. When using neural networks to predict PB samples were first screened for training, excluding the results of a small impact on the network output sample points.

Samples streamlined. Elementary row transformation matrix to maintain a linear relationship between the matrix column vector, this conclusion can be used to

[a] Corresponding author: zh1011@yxnu.net

streamline our sample data, sample data so streamlined to maintain a linear relationship between each attribute fields. Training the neural network is actually a given sample in real time to adjust the network connection weights through the process, the results of the sample pretreatment convergence for network training play a key role.

3 Analysis of practical examples

In this paper, using MIT Lincoln Laboratory test data KDDCUP99, it is specifically used for intrusion detection evaluation. We use a subset of the data set on 10% as the experimental data source, which contains a total of 494,021 network connections, which normally connected 97277, abnormal connections 396,744.

Here we analysis for data of DOS attack types (DOS attack type coded as "0001").

Let A is sample data matrix where each row vector representation of a sample data, the row vector containing 34 data, assuming that there is a records sample, then A is a matrix of 34 rows; due to enter the network after each sample data will have a corresponding output, coupled with the examples of the type of DOS attack with code "0001" indicates, the desired output matrix B is a matrix of a rows and 4 columns (temporary threshold is not considered, only consider the weight issue). In this paper, BP neural network input nodes is 34, the hidden layer nodes is 15, the output layer nodes for four although the connection weights neural network weights can be represented as a real number string, but during the training of the network, also need to be a real number string into two parts, set the input layer to the hidden layer connection weights matrix W1, then W1 is the matrix of 34 rows and 15 columns; Similarly, the hidden layer to the output layer connection weights weight matrix W2 is 15 rows and 4 columns. So we can get the formula (1).

$$A_{a\times 34}W^{1}_{34\times 15}W^{2}_{15\times 4} = B_{a\times 4} \quad (1)$$

In formula (2), A and B is the coefficient matrix, C is the augmented matrix. After elementary row after conversion formula (3) as shown with constraints.

$$C = \begin{bmatrix} A & B \end{bmatrix} \quad (2)$$

$$C = [A, B] = \begin{bmatrix} A' & B' \\ E & D \end{bmatrix} \quad (3)$$

In formula (3), C matrix and D matrix is 0, after subsequent processing, Corresponding to the output from the original A, B now becomes A ', corresponding to the output B', through this process, we can be a large sample into smaller sample, making computing faster, more streamlined sample data. In order to make the sample used in classification sync detector detects the proposed model, we first classify the sample data consolidation, were constructed DOS, PROBE, U2R, R2L four categories attack sample data sets, so that each of the four detectors detection four categories attacks. In order to reduce the number of suspicious attack, that attack the four types of data sets overlap between the number of records to be small. Increase the accuracy of calculation of the amount over the General Assembly, which will reduce the learning speed; precision is too small, will increase the number of records overlap, resulting in a suspicious number of attacks increases, the impact of training results.

Table 1 decimal reserved a table when recording duplication

DOS	44	0	37	4
PROBE	49	37	0	5
R2L	33	4	5	0
U2R	11	1	2	6
	Normal	DOS	PROBE	R2L

Table 2 decimal reserved four recording duplication

DOS	1	0	0	0
PROBE	1	0	0	0
R2L	0	0	0	0
U2R	0	0	0	0
	Normal	DOS	PROBE	R2L

Table 3 sample data compression case

	Before compression	**Decimal point one**	**Decimal point four**
Normal	97277	3918	85811
DOS	391459	228	37436
PROBE	4107	233	1775
R2L	1119	52	921
U2R	59	20	52

After optimization of the training samples of using the above method, using matrix elementary row transformation matrix to maintain a linear relationship between the column vectors this conclusion, we can further streamline the sample data, the sample data can be maintained so streamlined between the attribute fields linear relationship. Training the neural network is actually a given sample in real time to adjust the network connection weights through the process, the results of the sample pretreatment convergence for network training play a key role.

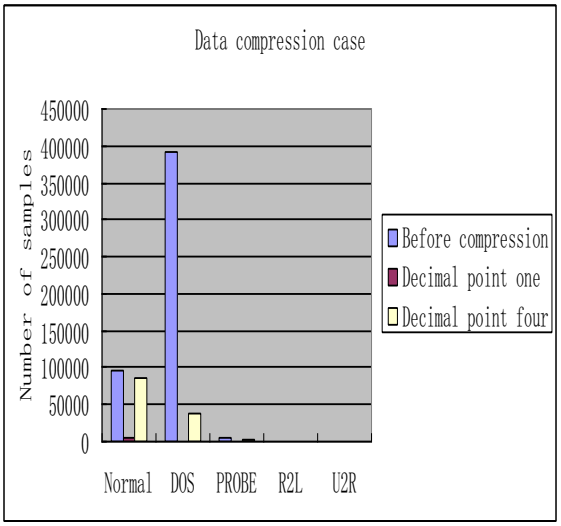

Figure 1 data compression case of FIG.

4 Conclusion

(1) Analysis of neural network nonlinear predictive multivariable predictive superiority and neural network for forecasting disadvantage lies.

(2) The proposed five important parameters affecting network forecasting capabilities. The sample quality, sample normalization, the input layer nodes, hidden nodes and the target value of MSE for network training tolerance.

(3) Training error allowed under certain circumstances, to study the impact on the network without a parameter prediction accuracy, and found there is an optimal sample input layer nodes and hidden nodes, such networks have strong predictive capacity.

(4) This paper is constructed using genetic algorithm neural network while optimizing impact prediction accuracy of the parameters (number of nodes in the input layer, hidden layer nodes and sample allows the training error) algorithm, has been on better network prediction model Finally, numerical examples verify the correctness of the analysis results.

References

[1] Chen Guo analyzed factors predictive accuracy of neural network model [J] Beijing: Pattern Recognition and Artificial Intelligence 2005,18 (5): 528-533

[2] Jiang Lin, Chen Tao, Qu Liangsheng affect the quality of the training sample performance of artificial neural networks [J] Beijing: China Mechanical Engineering 1979,8 (2): 50-53

Binding [3] LI Min-qiang, Xu Bo Yi, Ji-song Kou genetic algorithms and neural network [J] Beijing: System Engineering Theory and Practice 1999 (2):

[4] WU Huai-yu, SONG Yu-chieh neural network nonlinear regression analysis [J] Wuhan: Wuhan University of Science and Technology Metallurgy, 1998,21 (1): 90-93

[5] WANG Yi-huai, Wang Lin. Nonlinear regression based on artificial neural networks [J] Beijing: Computer Engineering and Applications 2004, 12, 79-81

[6] high-Hong Shen, SUPPORTED nonlinear regression artificial neural network prediction model [J] Beijing-based: North China University of Technology, 1999,11 (1), 68-73