IAOS Satellite Meeting on Statistics for the
Information Society
August 30 and 31, 2001,Tokyo, Japan

Relationship between Regression with qualitative variables and Feed-forward Neural Networks

ASANO, Miyoko
DAITO BUNKA University, Japan

Abstract

This paper reports the results of a comparison of application results between Feed-forward Neural Networks and Regression Analysis with qualitative variables. Neural Network, with relatively few hidden layer units, can provide concrete examples of the probability in taking into account classified variables in creating hierarchical structures. The given situation is the measurement of the momentary flux of water in restrooms of an office building. The input variables include classified factors such as: seasons and genders. By applying appropriate supportive variables to the Neural Network, without direct input of any classified factors, the effect of the factors are automatically accounted for and prediction can be made with substantial accuracy in Neural Networks models.

1 Introduction

Neural Network Analysis can be applied broadly to many fields. Its mathematical relationship to statistics has been reported and its ability to measure function approximation was concluded in summary reports by Bishop (1995), Ripley (1996). In feed-forward Neural Network, any types of function can be estimated if the unit number of the hidden layer is increased as according to Diaconis, P & Shahshani (1984). While Asoh (2001) discussed the relationship between statistics, information theory, statistical dynamics, and information geometry. While there is statistical discussion about Neural Networks Analysis, particularly, its practical application to data analysis, many theoretical factors are unclear. Although when regarding Neural Networks as a non-parametric regression model, it is useful to think about its potential in being applied to other statistical models. However, application is difficult because there are an abundant of expressions possible in Neural Networks models.
This paper presents data based on real situational conditions which indicates that Neural Networks models can express discontinuous regression including hierarchical factors with comparatively few number of units.

2 Comparison of Feed-Forward Neural Networks to Regression Models with Qualitative Variables

2.1 Effective prediction of the momentary volume of flowing water

Ito at al. (1999) applied this Neural Network data analysis to the momentary volume of flowing water. The precise prediction of momentary volume of flowing water is necessary when evaluating the use of water efficiently in the construction of a building. The momentary volume of flowing water was analyzed by using a 3-layered Feed-forward Neural Network. This model leaves room for discussion regarding the actual factors, which have influenced the amount of water used, like seasons or sex, which are not included in the variables studied. Therefore, in this study a comparison was done on the analytic precision between cases when season or gender were not included as variables as opposed to when they were included. The precision of prediction in Linear regression analysis models which included such variables as seasons or sex, were not as good as the Neural Network analysis. This

indicates the following:

My conjecture: By inputting appropriate variables, Neural Network analysis will automatically take potential effective factors (not yet inputted) into consideration.

To test this conjecture, we analyzed the relationship between the effects of a number of potential factors in the hidden layers.

2.2 **Data Analysis and the purpose of the data**

The input variables in predicting the momentary volume of flowing water are: the volume of flowing water per minute, per 10 hours, per 24 hours, and the maximum volume per one hour. The output variable is the maximum volume of flowing water per second. This data enables us to compare the result of regression analysis versus Neural Network analysis.

Table1. The variables of estimate the maximum flow rate

No.	variables	content of variables
1	X1	The maximum value in a minute
2	X2	The maximum value in an hour
3	X3	The sum in 10 hours
4	X4	The sum in a day
5	Y	The estimated maximum values in a second the determined values after estimation of the population distribution with sampled values, which were the 95%
6	Season	qualitative variables for Season : 2 levels
7	Genders1 Genders2	qualitative variables for Genders : 3 levels
8	Number	an appropriate auxiliary variable of qualitative variable

The following takes into account the use of layered Feed-forward Neural Networks and its significance: Seasons, and gender-1 and gender-2 have been included as possible factor variables. There were only 2 seasons-summer; indicated as: 'season=0', and winter-indicated as 'season=1'. Gender was divided into 3 categories: (1) men's restrooms, indicated as 'gender1=gender2=0' (2) women's restrooms, indicated as: 'gender1=0; gender2=1' and (3) other situations, indicated as: 'gender1= 1, gender2= 0'. In addition, serial numbers are used for each potential layered factor variable. Three-layered Feed-Forward Neural Networks were used for comparison, and the Sigmoid function was used for the output of the hidden layer. The linear function was used for the output of the final layer.

Table2. Comparison between Regression Analysis and feed-forward neural networks.

	Neual networks			regresion model		
input variable/dependent variables	units	R S S	A I C	multiple correlation coefficient	R S S	A I C
X4	9	2.03	-410.4	0.184	5.27	-352.5
X3,X4	5	1.93	-430.2	0.509	4.04	-381.0
X2,X3,X4	7	1.22	-452.5	0.518	3.99	-380.6
X1,X3,X4	4	1.76	-440.9	0.518	3.99	-380.5
X1,X2,X3,X4	11	0.62	-468.8	0.524	3.95	-379.6

Table3. Qualitative variabe Season plus input/dependent variables

	Neual networks			regresion model		
input variable/dependent variables	units	R S S	A I C	multiple correlation coefficient	R S S	A I C
X4,Season	6	2.51	-403.6	0.205	5.22	-351.5
X3,X4,Season	7	1.36	-440.5	0.509	4.04	-379.1
X2,X3,X4,Season	25	0.08	-529.5	0.519	3.98	-378.7
X1,X3,X4,Season	10	0.80	-451.3	0.518	3.99	-378.6
X1,X2,X3,X4, Season	24	0.01	-737.1	0.525	3.95	-377.7

To calculate the weight, we used the S-PLUS, NNET procedure and to measure the effect of the potential factor, we used real examples. Also taken into consideration were the effects of the input factor. The data analyzed was the volume of flowing water per second in the men and women's restrooms, restrooms for the disabled, and sinks (on the 20th and 36th floors). There were 36 points analyzed which were divided into 3 categories-full volume, moderate, and hot water) according to their sources. The period of observation was from Dec. 17 to 20 and also, from the 24th in 1996, from Aug. 25 to 29 in

1997. As per the explanation on Table 1, 5 variables were set-up according to: the volume of flowing water per minute, per 10 hours, per 24 hours, and the maximum volume per hour, and the predicted maximum volume of flowing water per second. There were 23 points observed. The number of days observed were: 5 days in the summer and 5 days in the winter, for a total of 115. The output variables are the predicted maximum volumes of flowing water per second, which is termed as "the 'maximum flux".

2.3 **Results and Considerations**

Table 2 to Table 6 compare the accuracy of prediction between data compiled from Neural Networks and Regression Analysis with qualitative variables. At the left of each table are the results of Neural Network and at the right are the results of Regression Analysis. The criteria for this data is the minimum AIC (Akaike (1974)), the number of units depending upon the variables of the hidden layer. The information criteria, MDL (Rissansan (1983)), uses the same models. The application of the AIC to 3-layered Feed-Forward Neural Networks followed the same procedures followed by Kurita in 1990. Table 2 is the results of combing defined factors with the exception of classified factors. This shows that when the number of factors increases that the liner regression analysis of AIC is improved very little. On the other hand, through Neural Network Analysis, AIC is very improved with the increase of the number of input variables. Table 3 shows the results of adding one classified factor, season to the results of Table 2. Table 3 shows that the additional season, one classified factor, does not improve the accuracy of the linear regression analysis. On the other hand, in Neural Networks, the addition of one season to the 24-hour factor does not improve the RSS or AIC, but there are some combination of input factors which do improve. It must be noted that there is an increase in the number of units in the hidden layers. Table 4 shows the results of adding Gender 1 and 2 to defined factors, which indicate improvement in the multiple correlation coefficients, RSS, and accurate prediction in linear regression analysis also improves as a result of the Neural

Table 4. Qualitative variable Gender plus input/dependent variables

	Neual networks			regresion model		
input variable/dependent variables	units	R S S	A I C	multiple correlation coefficient	R S S	A I C
X4,Gender1,Gender2	1	2.20	-444.8	0.752	2.37	-440.5
X3,X4,Gender1,Gender2	7	0.84	-482.2	0.788	2.07	-454.1
X2,X3,X4,Gender1,Gender2	10	0.60	-485.1	0.791	2.04	-453.7
X1,X3,X4,Gender1,Gender2	7	0.60	-506.3	0.789	2.06	-452.6
X1,X2,X3,X4, Gender1,Gender2	17	0.08	-564.2	0.792	2.04	-451.9

Table5. Qualitative variable Season & Gender plus input/dependent variables

	Neual networks			regresion model		
input variable/dependent variables	units	R S S	A I C	multiple correlation coefficient	R S S	A I C
X4,Season,Gender1,Gender2	1	2.08	-449.2	0.759	2.31	-441.4
X3,X4,Season,Gender1,Gender2	5	0.81	-500.1	0.791	2.04	-453.7
X2,X3,X4,Season,Gender1,Gender2	21	0.14	-516.7	0.794	2.01	-453.3
X1,X3,X4,Season,Gender1,Gender2	11	0.21	-548.8	0.792	2.03	-452.1
X1,X2,X3,X4,Season,Gender1,Gender2	31	0.00027	-934.7	0.795	2.01	-451.4

Table6. an appropriate auxiliary variable of qualitative variable plus input/dependent variables

	Neual networks			regresion model		
input variable/dependent variables	units	R S S	A I C	multiple correlation coefficient	R S S	A I C
X4,Number	13	2.23	-375.5	0.382	4.66	-366.7
X3,X4,Number	7	1.60	-435.5	0.565	3.71	-390.8
X2,X3,X4,Number	8	1.04	-461.0	0.568	3.69	-389.4
X1,X3,X4,Number	11	0.65	-463.8	0.568	3.69	-387.4
X1,X2,X3,X4,Number	8	0.72	-471.8	0.576	3.64	-395.1

Networks. So gender is an important factor to improving accuracy. When the number of variables in the hidden layer increases in comparison to the results in Table 2, except for in the 24-hour case, and compared to Table 3, it decreases. Table 5 shows the results when a season is added, and gender 1 and gender 2 are combined with other variables.

These Neural Network results are more accurate and precise than that of Table 3 and 4. Therefore, it proves that the Neural Network will indicate changes not apparent in linear regression analysis models. Table 6 shows what happens when serial numbers have been added to combinations of variables. Serial numbers are alternative variables, which are related to the input variables. When used, the number of the hidden layers decrease when 4 input variables are applied. When there are 4 input variables, the effective factors of the serial numbers are included, and there are fewer units. However, in other combinations, the correlation between the serial numbers with the effective factors are automatically taken into consideration, but the number of the hidden layer units are also increased. The result of regression analysis shows no improvement.

3 Conclusion

To compare Neural Networks with Linear Regression Analysis with qualitative variables, the following can be said: In the case of Regression Analysis, multiple classified factors and known variables need to be added to defined conditions.

(1) If there are any structural changes in the Neural Network, by applying appropriate input variables to the number of hidden-layer units, structural changes can be taken into consideration and analyzed.

(2) Adding alternative variables, or by clarifying input values, the effects of structural changes are automatically taken into account. If there are no classified factors, such as input variables, models include hierarchical factors with comparatively few units. The serial number is considered as an alternative variable and it plays the role of automatically taking into consideration the factors to give the Neural Networks model its effectiveness.

These characteristics of Neural Networks are not only more effective than regression analysis with potential factors, but also prove to be effective in determining solutions to specific problems.

Reference

Akaike (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19, 6,716-723.

Asano,Y.Asano,M.N.Ichikawa (2000). A Study the Estimation Method of the Maximum Load of Water Supply System in an Office Building by the Neural Network Model. *Technical Proceedings of the CIB W62 2000-26th International Symposium Rio De Janeiro on Water Supply and Drainage for Buildings.*

Bishop (1995). *Neural Networks for Pattern Recognition*, OXFORD UNIVERSITY PRESS.

Diaconis, P. Shahshahani, D. (1984). On non-linear function of linear combination. *SIAM Journal on Scientific and Statistical Computing* 5,175-191.

Ripley (1996).*Pattern Recognition and Neural Networks*. CAMBRIDGE UNIVERSITY PRESS.

Rissansan, J. (1983) A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11,2,416-431.

Asoh H.(2001). Information-Based Induction Sciences. *Journal of Japanese Society for Artificial Intelligence*, 16,2,287-299. (in Japanese)

Itou,h. Ichikawa,N. Yamauchi,O, Asano,Y.Baba,T,Asano,M. (1999). A Study the Estimation Method of the Maximum Load of Water Supply System in an Office Building by the Neural Network Model. *Journal of Japan Water Works Association*, 67 ,7,28-36. (in Japanese)

Kurita,T.(1990). A Method to Detective the Number of Hidden Units of Three Layered Neural

Networks by Information Criteria.*Journal of IEICE* J73-D-**II**,11,1872-1878. (in Japanese)