The Preliminary Analysis of Artificial Neural Network Model for Prediction

Isao Inoue

1. Introduction

This paper is an attempt to evaluate the prediction ability of artificial neural networks (ANNs) by using data obtained from the 2006 NORC General Social Survey (GSS). ANNs have often been used in classification and survival prediction in many biomedical areas, and proved to be more powerful than traditional statistical methods such as logistic regression (see Ahmed, F. E. (2005) and Eftekhar, R., K. et. al. (2005)). The ANN employed in this study is the Neural Networks add-on module included in the SPSS Statistics 17.0.

2. Materials and Methods

The following six variables are selected from the 2006 NORC General Social Survey (GSS) database (see Davis, J. A. et. al. (2007)):

(1) WRKSTAT (labor force status)
Response categories:
1=Working fulltime
2=Working parttime
3=Temp not working
4=Unemployed, laid off
5=Retired
6=School
7=Keeping house
8=Other

(2) DEGREE (respondent’s highest degree)
Response categories:
0=Less than high school
1=High school
2=Junior college

3=Bachelor’s
4=Graduate

(3) INCOME (total family income)
Response categories:
1=Less than $1000
2=$1000 to 2999
3=$3000 to 3999
4=$4000 to 4999
5=$5000 to 5999
6=$6000 to 6999
7=$7000 to 7999
8=$8000 to 9999
9=$10000 to 14999
10=$15000 to 19999
11=$20000 to 24999
12=$25000 or more

(4) MARITAL (marital status)
Response categories:
1=married
2=widowed
3=divorced
4=separated
5=never married

(5) HAPPY (general happiness)
Response categories:
1=very happy
2=pretty happy
3=not too happy

(6) CHILDS (number of children)
Response categories:
0=none
1=one
2=two
3=three
4=four

5=five
6=six
7=seven
8=eight or more

Responses such as the following are treated as missing values and excluded from the present study:

(7) a. Don't know
b. No answer
c. Not applicable
d. Refused

Thus, we will use the remaining 2573 cases to study how well ANNs can predict the dependent variable HAPPY on the basis of five predictor variables, namely WRKSTAT, DEGREE, INCOME, MARITAL, CHILDS. To simplify the analysis of results, the two levels of the dependent variable, *very happy* and *pretty happy* are merged into the single level, *happy*, so that the dependent variable, HAPPY is dealt with as a dichotomous category.

The 2573 cases are randomly assigned to the three samples, namely the training sample (50%), the testing sample (20%), and the holdout sample (30%). The training sample is used to train the ANN, and the testing sample is used to prevent overtraining, in order to obtain better generalization performance. The holdout sample is used to assess the predictive ability of the ANN, because this sample is not used in the training session.

The ANN employed in all our experiments has the following architecture: the input layer contains 39 units, the hidden layer contains 7 units, and the output layer is comprised of two units. The 38 input units are assigned to each response category of five predictor variables, and the remaining one is used as a bias input unit. The two output units correspond to the HAPPY dichotomous category. The function used in the hidden units is hyperbolic tangent, and transforms the weighted sum of inputs to the range of (-1, 1), while the output layer uses the softmax function, transforming input vectors to vectors with the range of (0, 1).

3. Results

20 trial runs from different random initial conditions are carried out and the average percent of correct predictions is 86.5 with *SD*=1.34 and range (83.5-89.1). The average ROC area is 0.71 with *SD*=0.02 and range (0.63-0.75) and this level is

considered acceptable (see Hosmer, D. W. and S. Lemeshow (2000: 162)). However, almost all cases of *not-too-happy* category (approximately an average of 99%) are incorrectly predicted as belonging under *happy* category, while almost all *happy* cases are correctly predicted (approximately an average of 99.5%). This total failure of discrimination between the two categories may result from the big difference in numbers between *happy* and *not-too-happy* subjects in the training samples. In the training samples, the average number of *happy* subjects is 1117, while that of *not-too-happy* subjects is 167. If the ANNs adopt the strategy of regarding all the cases as *happy*, they can achieve a high level of performance, average correct prediction rate of 87% in the training samples. Thus, in the holdout samples, they can also achieve the average success rate of 86.5% by predicting almost all the subjects as *happy*.

All the predictor variables except CHILDS are found effective in discriminating between *happy* and *not-too-happy* subjects, because the results of chi-square tests of WRKSTAT, DEGREE, INCOME, MARITAL predictor variables are significant ($p<0.001$). Thus, the initial rough analysis of these results suggests that the possible remedy for this problem might be to balance the number of cases of dichotomous category in the training samples.

References

Ahmed, F. E. (2005) "Artificial Neural Networks for Diagnosis and Survival Prediction in Colon Cancer," *Molecular Cancer* 4: 29.

Davis, J. A. et. al. (2007) *General Social Surveys, 1972-2006 Cumulative Codebook*, National Opinion Research Center, University of Chicago.

Eftekhar, R., K. Mohammad, H. E. Ardebili, M. Ghodsi, and E. Ketabchi (2005) "Comparison of Artificial Neural Network and Logistic Regression Models for Prediction of Mortality in Head Trauma Based on Initial Clinical Data," *BMC Medical Informatics and Decision Making* 5: 3.

Gorman, R. P. and T. J. Sejnowski (1988) "Analysis of Hidden Units in Layered Network Trained to Classify Sonar Targets," *Neural Networks* 1, 75-89.

Hosmer, D. W. and S. Lemeshow (2000) *Applied Logistic Regression*, John Wiley & Sons, New York, N. Y.

Obuchowski, N., M. L. Lieber, and F. H. Wians Jr. (2004) "ROC Curves in Clinical Chemistry: Uses, Misuses, and Possible Solutions," *Clinical Chemistry* 50: 7, 1118-1125.