On the Effect of Training Data on Artificial Neural Network Models for Prediction

Isao Inoue

1. Introduction

In Inoue (2009), it is found that highly unbalanced proportion of categories in the training data leads ANNs to neglect cases of the minor category and incorrectly predict that almost all cases belong to the preponderant category. In this paper, I would like to study how increasing the proportion of the minor category affects the prediction ability of artificial neural networks (ANNs). The ANN employed in this study is the neural networks add-on module included in the *SPSS Statistics 17.0* and the data are taken from *Work Status with Attitudinal Variables* in Tabachnick and Fidell (2001).

2. The Dataset

The following 11 variables are used as independent variables:

(1) MARITAL (current marital status)
  1=Single
  2=Married
  3=Broken

(2) CHILDREN (presence of children)
  0=No
  1=Yes

(3) RELIGION (religious affiliation)
  1=None or other
  2=Catholic
  3=Protestant
  4=Jewish

(4) RACE (ethnic affiliation)
  1=White
  2=Non-white

(5) CONTROL (measure of control ideology; internal or external)
(6) ATTMAR (satisfaction with current marital status)
(7) ATTROLE (measure of conservative or liberal attitudes toward role of women)
(8) SEL (measure of deference accorded employment categories)
(9) ATTHOUSE (frequency of experiencing various favorable and unfavorable attitudes toward homemaking)
(10) AGE (chronological age in 5-year categories)
(11) EDUC (years of schooling)

The dependent variable is the following binary category assessing current attitude toward unemployed status:

(12) WORKSTAT
2=Role-satisfied housewives
3=Role-dissatisfied housewives

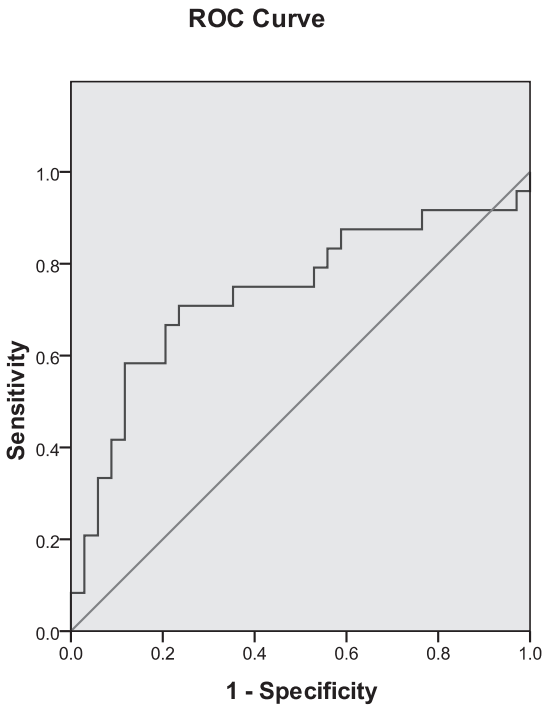
The dataset comprises 214 cases and is randomly partitioned into three samples, namely the training sample (49.5%), the testing sample (20.0%), and the holdout sample (30.5%). The training sample is used to train the ANN and the testing sample is used to prevent overtraining, in order to obtain better generalization performance. By using these two samples, the ANN acquires the knowledge of distributional pattern of the dataset. After the training session is completed, the holdout sample is used to assess the predictive ability of the ANN.

The ANN employed in this study has the following architecture: the input layer contains 42 units, the hidden layer contains 1 unit, and the output layer is composed of two units. The activation function of the hidden unit is hyperbolic tangent, transforming the weighted sum of inputs to the range of (-1, +1), while the output layer uses the softmax function, generating input vectors within the range of (0, 1).

3. Results

15 trial runs from different random initial conditions are carried out with the average proportion of the major category (role-satisfied housewives) to the minor category (role-dissatisfied housewives) being 63.2 to 36.8 in the training data, contrasting with the proportion of 85 to 15 adopted in Inoue (2009). After the training session, the holdout sample is applied to the ANN, which gives us the following results. The change in the proportion of the training data results in the improvement of the prediction ability, increasing the average percentage of correct predictions of

the minor category from 1% observed in Inoue (2009) to 46.2%, while the average percentage of correct predictions of the major category attains 77.0%. The average area under the ROC curve, which measures the ability of the ANN to correctly classify the categories, earns the level of 0.73 with *SD*=0.05 and range (0.62-0.82) and this level can be considered to be acceptable and fair. The following is the typical ROC curve with the area under the ROC curve of 0.737 we obtained during the predictability assessment.



The diagonal line represents completely random guesses and corresponds to the area under the ROC curve of 0.5, while the points above the diagonal line can be considered to be good classification results.

When we change the average proportion of the major category (role-satisfied housewives) to the minor category (role-dissatisfied housewives) in the training data to 82.5 vs 17.5, in contrast to the initial proportion of 63.2 vs 36.8, the average percentage of correct predictions of the minor category drops from 46.2% to 10.4%, while the average percentage of correct predictions of the major category attains

93.8%. These results and similar observations in Inoue (2009) show that the predictive ability of ANNs declines significantly with respect to the discrimination of dichotomous categories when the proportion of two categories become highly unbalanced.

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.