## Chapter IX

# Predicting Consumer Situational Choice with Neural Networks

Michael Y. Hu, Kent State University, USA

Murali Shanker, Kent State University, USA

Ming S. Hung, Optimal Solutions Technologies, Inc., USA

## ABSTRACT

*This study shows how neural networks can be used to model posterior probabilities of consumer choice and a backward elimination procedure can be implemented for feature selection in neural networks. Two separate samples of consumer choice situations were selected from a large consumer panel maintained by AT&T. Our findings support the appropriateness of using neural networks for these two purposes.*

# INTRODUCTION

In recent years, there has been an upsurge in the business applications of artificial neural networks (ANNs) to forecasting and classification. Examples include prediction of bank bankruptcies (Tam & Kiang, 1992), success in joint ventures (Hu et al., 1996, 1999a), consumer choices (Kumar et al., 1995; West et al., 1997), derivative/option, stock prices (Lo, 1996; Refenes et al., 1996), and forecasting of currency exchange rates (Hu et al., 1999b), to name a few. An extensive review of forecasting models using ANNs is provided in Zhang et al. (1998). Despite this upsurge, many market researchers still treat ANNs as black boxes. However, just like any statistical model, neural networks must be carefully modeled for the application to be successful. In this study, we consider the various aspects of building neural network models for forecasting consumer choice. Specifically, a situational consumer choice model is constructed, and neural networks are used to predict what product or service a consumer will choose. Our approach relies on the estimation of *posterior probabilities* for consumer choice. The posterior probability, being a continuous variable, allows more interesting analysis of the relationships between consumer choice and the predictor variables.

The type of ANNs that we consider are multi-layer feedforward networks. Probably the most popular training method for such networks is back-propagation (Rumelhart et al., 1986). In this study, we use the algorithm developed by Ahn (1996) for training. As feedforward networks are now well established and discussions can be found in most textbooks on neural networks, they will not be presented here. But, one frequent and valid criticism of neural networks is that they can not explain the relationships among variables. Indeed, since neural networks usually use nonlinear functions, it is very difficult, if possible at all, to write out the algebraic relationship between a dependent and independent variable. Therefore, traditional statistical relationship tests — on regression parameters, for example — are either impossible or meaningless. A typical approach in neural network modeling is to consider the entire network as a function and just investigate the predicted value of a dependent variable against the independent variables. In this chapter, such analysis is reported. In addition, we highlight two modeling issues when using neural networks:

- *Model selection.* Selection of an appropriate model is a nontrivial task. One must balance *model bias* (accuracy) and *model variance* (consistency). A more complex model tends to offer smaller bias (greater accuracy), but also greater variance (less consistency). Among neural networks, a larger network tends to fit a training data set better, but perform more poorly when it is applied to new data.

- *Feature selection.* A modeler strives to achieve parsimony. So the goal here is to build a model with the least number of independent variables, yet producing equal or comparable predictive power. For neural networks, as mentioned above, parameter testing does not apply and, therefore, more computational intensive methods must be employed to determine the variables that should be included in a model. We offer and validate a heuristic that has worked well for the test data set.

The organization of this chapter is as follows. The next section briefly discusses the use of posterior probabilities for consumer choice. The entire approach to model situational choice prediction is then illustrated. The data came from a large-scale study conducted by the American Telephone and Telegraph Company (AT&T). As can be expected, one of the objectives of the study was to find out how consumers choose between various modes of communication. The results are then presented, which is followed by a section describing an experiment that evaluates and validates our feature selection heuristic. The final section contains the conclusion.

# ESTIMATION OF POSTERIOR PROBABILITIES

## Definitions

A classification problem deals with assigning an object, based on its attributes, to one of several groups. Let **x** be the attribute vector of an object and $\omega_j$ denote the fact that the object is a member of group $j$. Then the probability $P(\omega_j \mid \mathbf{x})$ is called the posterior probability and it measures the probability that an object with attributes **x** belongs to group $j$. Traditional classification theory computes the posterior probability with the Bayes formula, which uses the prior probability and conditional density function (see, for example, Duda & Hart, 1973).

Posterior probabilities correspond to the likelihood of a consumer making a purchase in a consumer choice problem. Armed with the estimates of these probabilities, a marketer would know how likely a consumer is to alter his choice decision. For instance, a consumer with a probability of 0.498 is more likely to change one's choice than another with a probability of 0.20. Under this scenario, the marketer can more effectively target his product or messages to those consumers whose probabilities are closer to 0.5; and design strategies to increase these posterior probabilities for his product.

Typically, the posterior probability is a nonlinear function of **x** and cannot be derived directly. Hung et al. (1996) showed that the least squares estimators produce unbiased estimates of this probability. Neural networks provide a convenient way to perform this computation. For a prediction problem with ***d*** features and ***m*** groups, the neural network structure will have ***d*** input nodes and ***m*** output nodes. If we define the target values for the output nodes as in (1), and use the *least-square* objective function, it turns out the predicted value of the $j^{th}$ output variable is an unbiased estimator of the posterior probability that **x** belongs to group $j$ (Hu et al., 2000). For a two group prediction, only one output node is sufficient and the target values will be 1 for group 1 and 0 for group 2.

$$T_j^p = \begin{cases} 1 \text{ if object } x \text{ belongs to group } j \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

Two critical conditions must be met for the estimates of posterior probabilities to be accurate. One is sample size. In a previous study with simulated data sets (Hung et al., 1996), we found that the larger the training sample is, the greater the accuracy. The second is the network size. Theoretically speaking, the larger the network is (with more hidden nodes), the greater the accuracy of function approximation. However, for a given training sample, too large a network may lead to overfitting the sample, at the expense of generalization to the entire population.

## Model Selection

Model selection addresses the issue of what is the appropriate model (in our case, the neural network) for a given sample. Theoretically, model selection should be based on the trade-off between *model bias* and *model variance* (Geman et al., 1992). The bias of a model relates to the predictive accuracy of the model, whereas variance refers to the variability of the predictions. A model with low bias — by having many hidden nodes, for example — tends to have high variance. On the other hand, a model with low variance tends to have high bias. For a more detailed explanation of this issue, see Bishop (1995).

Empirically, we wish to select the smallest (in terms of hidden nodes) network with the best generalizability. A typical method to determine the generalizability of a model is to use a data set separate from the training set. In this project, the data set is divided into three subsets: *training*, *validation*, and *test sets*. For a given network architecture (here, it refers to the network with

a specific number of hidden and input nodes), the training set was used to determine the network parameters. The resultant network is then used to predict the outcome of the validation set. The architecture with the best generalizability is then chosen. The test set is used to measure how well the chosen model can predict new, unseen observations.

# EMPIRICAL EXAMINATION OF SITUATIONAL INFLUENCES ON CHOICE

The American Telephone and Telegraph Company maintained a consumer diary panel to study the consumer choice behavior in selecting long distance communication modes over time (Lee et al., 2000). The company embarked on a major research effort to understand the effect of situational influences on consumer choices of communication modes. It is envisioned that the usage of long distance phone calling is largely situational, since the service is readily available within a household and is relatively inexpensive. A demographically proportional national sample of 3,990 heads of households participated over a 12-month period. The sample was balanced with respect to income, marital status, age, gender, population density and geographic region. Each participant has to record the specifics on a weekly basis of one long distance (50 miles or more) communication situation. As a result, the company has compiled information on a total of roughly 250,000 communication situations.

## Choice Modeling

The communication modes being reported are of three types, long distance telephone calling (LD), letter or card writing. Since long distance telephone calling is verbal and the other two are nonverbal, letter and card in this study are combined into one category. The dependent variable, COMMTYPE, is coded as '1' for LD and '0' for 'letter and card.'

For a communication initiated by the consumer, information on five situation-related factors is also reported. These factors are:

- the nature (TYCALL) of the communication decision, whether it is 'impulse' (coded as '0') or 'planned' (coded as '1');
- reasons (REASON) for communication, 'ordinary' (coded as '1') or 'emergency' (coded as '0');
- receivers (RECEIVER) of the communication, 'relatives' (coded as '1') or 'friends' (coded as '0');

- total number of communications made and received (TOTALCOM) during the diary week; and
- total number of LD calls made and received (NUMCALLS) during the diary week.

Information gathered on TYCALL, REASON, and RECEIVER has marketing implications for how the long distance call services can be positioned in an advertising campaign. Also, based on past studies, the company has found that as TOTALCOM increases for a consumer, the frequency of using LD increases. Thus, a viable strategy is to remind a consumer to keep in touch with friends/relatives. Information on NUMCALLS also has implication for advertising positioning. Consumers, in general, tend to reciprocate in their communication behavior. When a phone call is received, a consumer is likely to respond by calling. The company can encourage consumers to respond when a call is received.

In addition, information on six consumer demographic and socioeconomic variables is also reported at the start of the diary keeping activities. These variables include number of times the consumer has moved his/her place of residence in the past five years (MOVES); number of relatives (RELATIVE) and friends (FRIENDS) that live over 50 miles or more away; age (AGE), average number of cards and letters sent in a typical month (NUMCLET) and average number of long distance telephone calls made in a typical month (MEANCALL).

In this study, we use all five situation-based and the six demographic variables to predict choice of modes. These demographic variables are potential candidates for segmentation while allowing the differences in situational and demographic influences to be captured.

A sample of 1,480 communication situations is used from the weekly diary database, 705 (47.64%) are LD calls made and the remaining 775 (52.46%) written communications. The entire sample of situations is from a total of 707 diarists. The maximum number of situations reported is three per diarist.

For neural network modeling, the data set, as mentioned before, is randomly partitioned into training, validation, and test sets. The distribution is 60%, 20%, 20% — exactly the same as in West et al. (1997). The specific composition is shown in Table 1.

## Design of Neural Network Models

As previously mentioned, the networks used in this study are feedforward networks with one hidden layer. There are direct connections from the input

*Table 1: Number of Situations by Choice and Partition*

| | Training | Validation | Test |
|---|---|---|---|
| Long Distance Call | 440 | 134 | 131 |
| Letter/Card | 448 | 162 | 165 |
| Total | 888 | 296 | 296 |

layer to the output layer. There is one output node and only it has a scalar. The activation function of the hidden nodes and the output node is logistic. An issue in neural network modeling is the scaling of input variables before training. Previous research (Shanker et al., 1996) indicates that data transformation is not very helpful for such problems and, hence, it is not performed here.

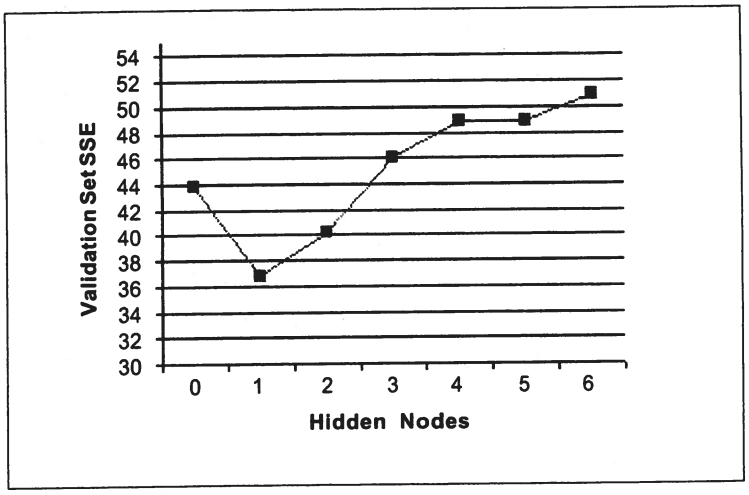
Given the choices made above, model selection is now reduced to the determination of the number of hidden nodes. Several practical guidelines have been proposed: ***d*** (Tang & Fishwick, 1993), ***2d*** (Wong, 1991), and ***2d+1*** (Lippmann, 1987), for a one-hidden-layer of ***d*** input nodes. However, none of these heuristics works well for all problems. Here we start with a network of 0 hidden nodes. It is trained on the training set and then applied to the validation set. Next we train a network of 1 hidden node and calculate the validation set *sums of square error* (SSE) similarly. This is repeated until a reasonably large number of hidden nodes has been investigated. (This number cannot be predetermined because the validation set SSE may go up and down for some time until a pattern develops.) Figure 1 shows the plot of SSE for the validation set as the number of hidden nodes varies from 0 to 6, with all the eleven feature variables as inputs.

Since the SSE in the validation sample takes on the smallest value at 1 hidden node, this architecture is selected for subsequent runs.

## Selection of Input Variables

As discussed earlier, feature selection is an important and difficult topic in neural network modeling. Since hypothesis tests on parameters are not applicable here, we resort to a backward elimination method. Train a network with all ***d*** features included. Then delete one variable and train a new network. Delete a different variable from the original set and train another new network. We end up with ***d*** networks, each having ***d-1*** features. Select the network with the smallest validation set SSE. Now consider the selected set of features as the

*Figure 1: Validation Set SSE vs. Number of Hidden Nodes*

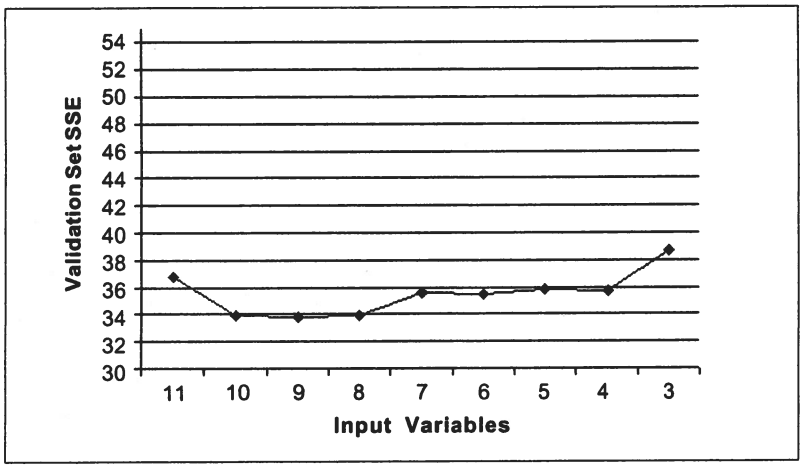



original set and repeat the process. This process continues until the validation set SSE increases drastically. This heuristic is admittedly "brute force," but the resultant network has been shown to classify better than the full-featured network in previous studies (Hu et al., 1996; Hung et al., 2001; Shanker, 1996). In addition, the following sections present an experiment that shows the backward-elimination approach for this dataset indeed selects the best of all possible network models.

As indicated in Figure 2, the validation set SSE for the 11-variable model is around 37. It drops to around 34 for the 10-, nine- and eight-variable models. It increases to about 36 and remains there for seven- to four-variable models. The next variable removal brings about a sharp increase in SSE. Although the eight-variable model has the smallest SSE, the four-variable is more attractive, because with only half of the variables its SSE is only slightly higher. So we decided on that model for further analysis.

The variables selected are REASON, RECEIVER, TOTALCOM and NUMCALLS. It is interesting to note that all the demographic variables are excluded from the final model. Researchers have found that situational and contextual factors have a major impact on situation-based choices (Hui & Bateson, 1991; Simonson & Winer, 1992). Conceptually, one can expect situation-specific factors to exercise a greater impact on these choices, since the consumer demographic factors are more enduring in nature and thus their influences may or may not enter into a particular purchase situation.

*Figure 2: SSE vs. Number of Features*




The appropriateness of the architecture being used is verified again by experimenting with the number of hidden nodes from 0 to 6. Once again the architecture with one hidden node is selected.

# RESULTS

## Investigation of Relationships

Suppose we knew that the four features — REASON, RECEIVER, TOTALCOM and NUMCALLS — would be useful to predict the type of communication. We can carry out some preliminary analyses before the models are built. Two of the variables, REASON and RECEIVER, are zero-one variables, so contingency tables such as Table 2 can be used.

Each ratio is the proportion of long distance calls with respect to the total number of communications. The observations are those in the training and validation sets. For example, there are 83 communications for REASON = 0 (emergency) and RECEIVER = 0 (friends), among them 59 are telephone calls. In general, the likelihood of placing a LD call is substantially higher in emergency situations and when the call is placed with relatives.

The other two variables are continuous and their relationships with the dependent variable can be explored with scatter plots. It was difficult to see any

*Table 2: Frequency Table for COMMTYPE*

| | REASON | | |
|---|---|---|---|
| RECEIVER | 0 | 1 | Total |
| 0 | 59/83 = .711 | 191/545 = .350 | 250/628 = .398 |
| 1 | 84/103 = .816 | 240/453 = .530 | 324/556 = .583 |
| Total | 143/186 = .769 | 431/998 = .432 | 574/1184 = .485 |

relationship between the dependent variable and either of the continuous variables in the scatter plots. In the interest of brevity, the plots are not shown here.

A neural network with one hidden node and four input nodes was trained on the combined training and validation sets, using the four features selected. The posterior probability is the probability that a consumer will choose long distance call for communication. So the first question a marketer may ask is, what is the relationship between each situational variable and such a choice? Table 3 shows the mean posterior probability for each combination of REASON and RECEIVER. The same pattern observed in the contingency table is clearly visible again — the probability to use long distance is highest under emergency situations to relatives. The fact that the average posterior probabilities are reasonably close to the raw relative frequencies in Table 2 confirms the validity of our neural network estimations.

With posterior probability as the dependent variable, and TOTALCOM and NUMCALLS as the two continuous independent variables, Figure 3 shows a clear pattern when REASON = 0 and RECEIVER = 0. First, the

*Table 3: Mean Posterior Probability*

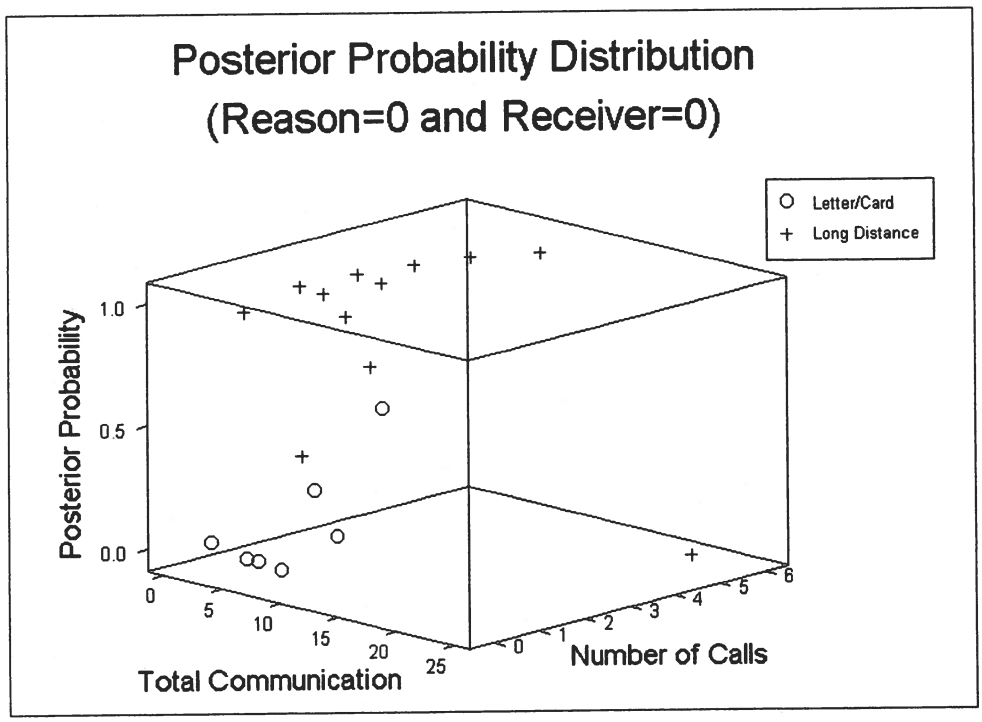
| | REASON | | |
|---|---|---|---|
| RECEIVER | 0 | 1 | Total |
| 0 | .744 | .323 | .379 |
| 1 | .781 | .514 | .564 |
| Total | .764 | .410 | .466 |

posterior probability functions are all nonlinear functions of TOTALCOM and NUMCALLS. Second, the function suggests a positive relationship with respect to NUMCALLS. With respect to TOTALCOM, the relationship is not clear when the variable is small, but seems positive when it is high. Similar patterns were observed in the other three plots and will not be presented here.

Some marketing implications can be drawn from the results of these graphs. The positive relationship between the posterior probability and NUMCALLS suggests that when a phone call is received, it is more likely for a consumer to respond with the same mode of communication. Notice that the process of reciprocity being generated can potentially lead to a multiplicative effect on the total volume of calls being made. A long distance phone company is well advised to remind consumers to reciprocate any long distance communication with the same mode.

Our results imply that as the total number of communication situations made and received (TOTALCOM) is small, the probability of making an LD call is widely scattered from 0 to 1; hence, it is difficult to predict the choice. However, when TOTALCOM is large (roughly more than 30), then the

*Figure 3: Posterior Probability Function*



probability of placing an LD call is very high, close to 1. In addition, as TOTALCOM goes up, the number of LD calls made should go up also. Therefore, it would benefit a long distance telephone company to encourage consumers to communicate more.

## Predictive Accuracy

To evaluate the ability of neural network models to generalize to previously unseen objects, a total of three models are constructed. The first includes all 11 original features. The second includes seven features selected by the backward elimination procedure in logistic regression (SAS, 1998). And the third uses only the four features selected by our own backward elimination procedure. For ease of reference, the lists of features are provided below.

- All 11 features: MOVES, RELATIVE, FRIENDS, AGE, NUMCLET, MEANCALL, TYCALL, REASON, RECEIVER, TOTALCOM, NUMCALLS.
- The seven features selected by logistic regression: NUMCLET, MEANCALL, TYCALL, REASON, RECEIVER, TOTALCOM, NUMCALLS.
- The four features selected by neural network: REASON, RECEIVER, TOTALCOM, NUMCALLS.

A neural network was built for each feature set and data used were the combined training and validation sets. The optimal number of hidden nodes for the seven-feature model was again one. Each feature set was also used to build a logistic regression model. All six models were then asked to predict the observations in the test set. Their performance is summarized in Table 4. The classification rate is based on the fact that there are a total of 296 observations in the test set, of which 131 involve long distance calls and the remaining 165 involve letters/cards.

Several important observations can be made. First, the neural network models are superior to logistic regression models in all cases except one (seven features, long distance). Second, the four-feature model outperforms every other model. This speaks voluminously for the merit of feature reduction used in this study. It also validates our own feature selection procedure. Third, the feature selection scheme for both neural networks and logistic regression seems able to find the optimal model: four-variable model for the former and seven-variable model for the latter.

The next section discusses model selection in greater detail, and presents experiments to validate our backward elimination feature-selection method.

# MODEL SELECTION

In all nonlinear models, including ANNs, model selection consists of specifying the nonlinearity component and feature selection. Architecture selection in ANN corresponds to specifying the nonlinear structure.

## Architecture Selection

Typically, the size of a neural network refers to its number of parameters (i.e., the number of arc weights and node biases). Given that we are concentrating on networks of one layer, the size of a network is directly related to the number of hidden nodes.

The methods to determine the appropriate network architecture can be summarized as follows:

1. Eliminate arcs whose weights are small or zero. Cottrell et al. (1995) construct an approximate confidence interval for each weight and if it contains zero, then the arc is eliminated.
2. Eliminate arcs whose *saliency*—a measure of relative importance—is small. Saliency is typically based on the partial derivative of the SSE with respect to the arc. Methods differ in the approximation of this derivative. The *optimal brain damage* of Le Cun et al. (1990) defines saliency of arc $i$ as $H_{ii}w_i^2/2$ where $H_{ii}$ is the i-th diagonal element of the *Hessian* matrix,

*Table 4: Classification Rates (Correct Classifications) for the Test Set*

| Model | Group | Neural Network | Logistic Regression |
|---|---|---|---|
| 11 Features | Total | .818 (242) | .787 (233) |
| | Long Distance | .870 (114) | .817 (107) |
| | Letter/Card | .776 (128) | .764 (126) |
| 7 Features | Total | .818 (242) | .804 (238) |
| | Long Distance | .763 (100) | .817 (107) |
| | Letter/Card | .861 (142) | .794 (131) |
| 4 Features | Total | .831 (246) | .794 (235) |
| | Long Distance | .840 (110) | .779 (102) |
| | Letter/Card | .824 (136) | .806 (133) |

the matrix of second derivatives (of SSE with respect to arc weights), and $w_i$ is the weight of arc $i$. The *optimal brain surgeon* (Hassibi & Stork, 1993), on the other hand, uses the diagonal element of the inverse of the Hessian matrix.

3. Build networks with different numbers of hidden nodes and then select one using some performance measure. The measure used by Moody and Joachim (1992) is called the *prediction risk* and it is the mean squared error on the validation set, adjusted by the number of weights. They also compute the prediction risk by using cross-validation, which first divides a data set into $k$ subsets and uses $k$-$1$ subsets for training and the $k^{th}$ subset for validation. The validation set then rotates to the first subset, and then to the second, etc., in a round-robin fashion.

As discussed in the next section, our paper uses a measure similar to that of Moody and Joachim (1992). For other methods, please see Bishop (1995, section 9.5).

## Feature Selection

In modeling, the principle of parsimony is important. Feature selection refers to the process of determining which subset of input variables is to be retained. It is a standard procedure in conventional pattern recognition (see, e.g., Fukunaga, 1990). Clearly one can use the methods mentioned above to eliminate one arc at a time until an input node is disconnected from the network and is thus eliminated by default. However, more efficient methods can be developed for this purpose.

There are two general approaches used in feature selection: *forward addition* and *backward elimination*. The former successively adds one variable at a time, starting with no variables, until no attractive candidate remains. The latter starts with all variables in the model and successively eliminates one at a time until only the "good"' ones are left. Whether a variable is attractive or not depends on its contribution to the model. For linear regression, well known measures for identifying good subsets of variables include (degree of freedom-adjusted) mean square error and prediction sum of squares (PRESS). For detailed discussions, see Neter et al. (1996) and Draper and Smith (1981).

In general, since backward elimination starts with the entire set of input variables, it is less likely to overlook any one variable's contribution in explaining the variability in the dependent variable, thus it is more likely for the procedure to arrive at the smallest subset of desirable variables.

For neural networks, several measures have also been proposed. Belue and Bauer (1995) calculate the (absolute) derivative of the *SSE* over each variable (called *saliency metric*) and drop the variables whose saliency is small. Moody and Joachim (1992) develop a sensitivity analysis (of a variable on *SSE*) based on their prediction risk and eliminate variables whose sensitivity is low. For other methods, please see Bishop (1995, section 8.5).

The next section presents our proposed method, which uses the backward elimination method for feature selection.

## Proposed Feature Selection Method

Our proposed method for feature selection is a backward elimination method based on our measure of prediction risk, which is very similar to that of Moody and Joachim (1992). Given a trained network of $n$ features and $h$ hidden nodes, denoted as $M^h_n$, the prediction risk is the mean sum of squared errors of a validation set $V$. That is:

$$MSE(M_n^h) = \frac{1}{|V|} SSE(M_n^h) = \frac{1}{|V|} \sum_{p=1}^{|V|} \sum_{j=1}^{l} (Y_j^p - T_j^p)^2$$

where $|V|$ is the number of patterns in the validation set $V=(Y,T)$ and $l$ is the number of output nodes of the neural network $M^h_n$. As the validation sets in our study are all of the same size, we use the sums of square error $SSE(M^h_n)$ as a measure of prediction risk in our method below.

1. Start with all $n$ features and train a network over a range of hidden nodes; i.e., $h = 0, 1, 2, \ldots$
2. Select the optimal hidden nodes $h^*$ which yields the smallest sums of square error $SSE(M^h_n)$.
3. Reduce the number of features by one, and train every possible $(n-1)$ feature network with $h^*$ hidden nodes. Let $SSE^*(M^{h^*}_{(n-1)})$ indicate the network with the smallest SSE of the $(n-1)$ networks.
4. If $(SSE^*(M^{h^*}_{(n-1)}) - SSE^*(M^{h^*}_n)) < D$, where $D$ is a predetermined positive quantity, then $n = (n-1)$, and go to Step 3. Otherwise, go to Step 5.
5. Use the features selected in Step 3, train networks over the range of hidden nodes used in Step 1 and select the optimal hidden nodes $h^*$ again.

## Experiment

An experiment was conducted to evaluate the backward elimination procedure. The experiment consists of training neural networks with all

possible combinations of the features of a data set and computes the prediction risks of each trained network. Results from the backward elimination procedure will then be compared with those from all possible combinations. The initial set of variables was the seven variables chosen by logistic regression.

A second random sample of 3,377 communication situations is drawn from the weekly diary database, where 1,594 (47.20%) entail LD calls and the remaining 1,783 (52.80%) involve written communications. The entire sample of situations is from a total of 2,111 diarists. The maximum number of situations is three per diarist. Since the primary objective is model selection, only training and validation samples will be needed. Of these 3,377 observations, 1,535 are used as training and the remaining 1,842 as validation. To measure the robustness of the backward elimination procedure, the validation sample is subdivided into three sets of equal size with Set 1 composed of 286 LDs and 328 written; Set 2 of 278 LDs and 336 written, and Set 3 of 298 LDs and 316 written. This cross validation scheme will show how sensitive model selection is with respect to validation samples.

## Network Architecture

All networks used have one output node, since there is one target variable COMMTYPE, and one hidden layer with $h$ hidden nodes. There are arcs connecting each input node to both the output node and the hidden nodes. The activation function at each hidden node and the output node is logistic. In addition, each hidden node has a scalar. For the purpose of model selection, the number of hidden nodes $h$ varies from 0 to 7.

## Results

A neural network was set up for each of the 127 possible combinations of the seven input variables. Each network was then trained using eight different architectures (zero to seven hidden nodes). These correspond to a total of 1,016 networks. Table 5 shows the minimum SSEs across all hidden nodes and sets of input variables for each validation sample. In Sample 1, among the seven one-variable networks, variable six (not shown) with four hidden nodes is tied with variable six with three hidden nodes with *SSE* equal to 103.87. Among the six-variable networks, the network with two hidden nodes has the minimum *SSE* of 68.62. The network with the smallest *SSE* among all combination of variables and hidden nodes is shown in bold.

Results from validation Set 2 are similar to those from Set 1. Both indicate that the six-variable network with variables 2, 3, 4, 5, 6 and 7, and two hidden nodes has the smallest *SSE*. Validation Set 3 shows a slight difference from the

other two samples. The four-variable (variable 4, 5, 6, 7) with two hidden nodes has the smallest *SSE*.

Next, we experiment with the backward elimination procedure. The seven input variables were trained in eight network architectures, hidden nodes from zero to seven. With validation sample 1, Table 5 shows that the network with two hidden nodes has the smallest SSE of 73.73 for seven variables. With the number of hidden nodes fixed at two, we then proceeded to examine the *SSE*s from the seven six-variable networks. As shown in Table 6, the network with variables 2, 3, 4, 5, 6, 7 has the smallest *SSE*, 68.62. Further elimination of variables resulted in an increase in *SSE*. The set of variables 2, 3, 4, 5, 6, 7 is then used to train networks of 0 to 7 hidden nodes, and the minimum *SSE* corresponds to the network with two hidden nodes (see Table 7). So the recommended feature set, based on validation sample 1, is (2, 3, 4, 5, 6, 7) and the network architecture is the one with two hidden nodes. This is the "best"' selection indicated by the all-combination experiment (Table 5).

With validation sample 2, the backward elimination method ends with the same "best" selection. The minimum *SSE* is 61.80. For validation sample 3, the backward elimination method starts with three hidden nodes for all seven variables and ends with four variables — 4, 5, 6, 7. Table 6 shows the *SSE* for this combination is 72.48. The set of four variables is then used to train networks of zero to seven hidden nodes and the minimum *SSE* corresponds to the network with two hidden nodes (see Table 7). This is the same as the best selection in the all-combination experiment (Table 5).

Overall results indicate that the backward elimination procedure identifies the same "best" models as the all-possible-combinations approach in each of the three validation samples. Neural networks are quite robust with respect to architecture and feature selection. Networks with two or three hidden nodes seem to be appropriate for this data set.

From a practical perspective, there seems to be little difference between models of six features and those of four features. In validation samples 1 and 2, the four-variable models end up with only a slight increase in *SSE* over the six-variable models. For example, in validation sample 1, the four variable model 4, 5, 6, 7 leads to an *SSE* of 70.82 compared to the smallest *SSE* of 68.62 for the six-variable model. However, a four-variable network with two hidden nodes has only 14 arcs, whereas a six-variable network with two hidden nodes has 20 arcs. A researcher can easily justify the selection of the four-variable model because of the greater reduction in the size of the network (which translates into greater degree of freedom for statistical analysis).

*Table 5: Minimum SSE Across Hidden Nodes and Number of Variables*

| # of Variables | Number of Hidden Nodes | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Validation Sample 1 | | | | | | | | |
| 1 | 114.68 | 106.13 | 106.13 | 103.87 | 103.87 | 115.04 | 114.74 | 115.24 |
| 2 | 101.40 | 84.45 | 77.78 | 78.81 | 79.54 | 80.27 | 81.80 | 80.83 |
| 3 | 98.74 | 79.82 | 73.72 | 74.70 | 76.30 | 77.31 | 77.48 | 76.72 |
| 4 | 95.45 | 76.91 | 70.82 | 71.54 | 73.03 | 73.18 | 73.74 | 73.97 |
| 5 | 92.88 | 74.38 | 68.68 | 70.23 | 69.95 | 73.18 | 74.66 | 75.45 |
| 6 | 92.24 | 75.37 | **68.62** | 70.73 | 72.37 | 72.88 | 73.32 | 75.29 |
| 7 | 92.29 | 75.51 | 73.73 | 74.38 | 77.65 | 78.31 | 80.84 | 82.72 |
| Validation Sample 2 | | | | | | | | |
| 1 | 115.19 | 103.11 | 103.11 | 98.27 | 98.27 | 110.73 | 109.94 | 110.01 |
| 2 | 87.17 | 80.58 | 69.54 | 70.37 | 70.17 | 70.86 | 71.76 | 72.37 |
| 3 | 86.21 | 79.44 | 67.70 | 68.09 | 68.66 | 70.25 | 70.47 | 70.85 |
| 4 | 83.27 | 75.63 | 64.50 | 65.06 | 66.24 | 67.17 | 67.31 | 68.06 |
| 5 | 82.74 | 74.29 | 63.19 | 64.78 | 64.98 | 66.51 | 69.43 | 70.18 |
| 6 | 82.88 | 73.63 | **61.80** | 63.87 | 64.25 | 64.63 | 65.93 | 66.79 |
| 7 | 83.14 | 73.67 | 66.46 | 67.73 | 71.31 | 74.24 | 74.65 | 75.46 |
| Validation Sample 3 | | | | | | | | |
| 1 | 118.07 | 108.24 | 108.24 | 108.17 | 108.17 | 111.93 | 111.89 | 112.19 |
| 2 | 96.29 | 84.18 | 75.00 | 75.19 | 75.74 | 76.64 | 76.51 | 76.97 |
| 3 | 94.76 | 83.90 | 75.08 | 74.04 | 75.62 | 74.89 | 75.04 | 77.15 |
| 4 | 91.91 | 79.41 | **72.06** | 72.48 | 72.74 | 73.20 | 74.67 | 75.80 |
| 5 | 91.26 | 78.85 | 73.11 | 73.23 | 72.66 | 75.55 | 76.11 | 78.29 |
| 6 | 91.52 | 79.74 | 74.03 | 75.55 | 76.09 | 75.21 | 77.68 | 77.04 |
| 7 | 91.73 | 80.57 | 76.80 | 76.13 | 78.08 | 78.10 | 78.66 | 80.14 |

# CONCLUSION

Applications of neural networks in marketing research are just now emerging. The few marketing studies we have identified all focused on using the technique for classification problems in particular choice decisions. Marketers are obviously interested in consumer choices. Prior researchers have shown the classification rates attained by neural networks to be superior to those by the traditional statistical procedures, such as logistic regression and discriminant analysis. Yet, marketers are also more interested in the likelihood of a choice outcome than the simple aggregate percentage of consumers choosing a product over the other.

Our study has shown that the posterior probabilities of choice can be estimated with neural networks via the least squares principle and that neural network in fact provides a direct estimate of these probabilities. Thus, the focus of this study is on the estimation of these posterior probabilities and the

*Table 6: Backward Elimination Procedure for All Validation Samples*

| Validation Sample 1 | | Validation Sample 2 | | Validation Sample 3 | |
|---|---|---|---|---|---|
| Variables Selected | SSE | Variables Selected | SSE | Variables Selected | SSE |
| 1234567 | 73.73 | 1234567 | 66.46 | 1234567 | 76.13 |
| Start with the above 7 variable model. | | | | | |
| 123456 | 89.44 | 123456 | 84.45 | 123456 | 95.78 |
| 123457 | 97.65 | 123457 | 89.43 | 123457 | 98.32 |
| 123467 | 71.71 | 123467 | 68.89 | 123467 | 80.13 |
| 123567 | 75.71 | 123567 | 68.16 | 123567 | 79.51 |
| 124567 | 77.02 | 124567 | 67.89 | 124567 | 76.97 |
| 134567 | 72.87 | 134567 | 64.91 | 134567 | 76.12 |
| **234567** | **68.62** | **234567** | **61.80** | **234567** | **75.55** |
| Use the best 6 variable model (shown in **bold** above). | | | | | |
| 23456 | 90.30 | 23456 | 91.57 | 23456 | 98.27 |
| 23457 | 97.53 | 23457 | 90.08 | 23457 | 97.47 |
| 23467 | 71.31 | 23467 | 67.68 | 23467 | 76.57 |
| 23567 | 71.51 | 23567 | 64.73 | 23567 | 76.45 |
| 24567 | 75.40 | 24567 | 65.36 | 24567 | 78.70 |
| **34567** | **68.68** | **34567** | **63.19** | **34567** | **73.23** |
| Use the best 5 variable model. | | | | | |
| 3456 | 91.50 | 3456 | 93.14 | 3456 | 97.27 |
| 3457 | 98.14 | 3457 | 93.40 | 3457 | 100.21 |
| 3467 | 70.98 | 3467 | 66.28 | 3467 | 75.30 |
| 3567 | 70.87 | **3567** | **64.50** | 3567 | 74.90 |
| **4567** | **70.82** | 4567 | 65.31 | **4567** | **72.48** |
| Use the best 4 variable model. | | | | | |
| 456 | 93.66 | 356 | 99.02 | 456 | 96.93 |
| 457 | 103.02 | 357 | 99.02 | 457 | 100.36 |
| **467** | **79.65** | **367** | **67.70** | **467** | **74.04** |
| 567 | 132.21 | 567 | 106.76 | 567 | 116.07 |
| Use the best 3 variable model. | | | | | |
| 46 | 97.94 | 36 | 100.87 | 46 | 100.72 |
| 47 | 108.73 | 37 | 105.91 | 47 | 107.14 |
| **67** | **77.78** | **67** | **69.54** | **67** | **75.19** |
| Use the best 2 variable model. | | | | | |
| **6** | **106.13** | **6** | **103.11** | **6** | **108.17** |
| 7 | 119.37 | 7 | 112.39 | 7 | 113.74 |

nonlinear functional relationships between these probabilities and the predictor variables.

Most market researchers treat neural networks as a black box. They leave the decision on model selection to computer software packages (if the packages have such capabilities) and typically rely on logistic regression for feature selection. Our study encompasses a rather comprehensive approach to neural network modeling. It provides guidelines for sample selection and shows how model selection should be carried out experimentally. A backward

*Table 7: SSE Across Hidden Nodes*

| Hidden Nodes | Validation Sample 1 | Validation Sample 2 | Validation Sample 3 |
|---|---|---|---|
| | Variables: 234567 | Variables: 234567 | Variables: 4567 |
| 0 | 92.24 | 82.88 | 91.91 |
| 1 | 75.37 | 73.63 | 79.41 |
| 2 | **68.62** | **61.80** | **72.06** |
| 3 | 70.73 | 63.87 | 72.48 |
| 4 | 72.37 | 64.25 | 72.74 |
| 5 | 72.88 | 64.63 | 73.20 |
| 6 | 73.32 | 65.93 | 76.39 |
| 7 | 75.29 | 66.79 | 76.28 |

elimination procedure adapted in this study actually identified a parsimonious model with even better classification rate. These results truly attest to the nonlinear modeling capabilities of neural networks.

The situational choice data set from AT&T contains variability over time and across consumers. Dasguta et al. (1994) report that most neural network applications have been with aggregate consumer data. There are only a handful of applications with disaggregate consumer survey response data. Data at a lower level of disaggregation typically contains more noise. Results reported in this study illustrate the potential for superior performance of neural networks for this domain of applications.

The variables retained by our feature selection procedure are all situation-based. As indicated in previous research in situational influences, situation-based factors should have a stronger bearing on situational choices as compared to the more enduring, consumer factors. This finding provides some validation for our suggested procedure. The nonlinear relationship between the posterior probabilities and the input variables was clearly captured graphically in our study. It is shown that these probabilities are more informative and useful for marketers in planning their strategies.

Practical managerial implications can be drawn from the results of this study. The benefits of long distance phone calling, particularly in emergency situations, are to be reinforced. Also, consumers are to be reminded that when communicating with relatives, long distance phone calling is the preferred choice. In addition, consumers are to be reminded to reciprocate in terms of modes of communications. When a consumer receives a long distance phone call, the consumer should be encouraged to use the same mode of communication in his/her response. Lastly, a long distance phone company should continuously remind its consumers to keep in touch with one's friends and

relatives. As the total frequency of communications increases, the likelihood of using long distance phone calling also goes up.

Major advances have been made in the past decade in neural networks. This study intends to introduce some of these major breakthroughs for researchers in the field of marketing. It is our hope that market researchers will be able to gain a better appreciation of the technique. Of course, these advances are available at a cost. Neural networks are much more computationally intensive than classical statistical methods such as logistic regression. The model selection and feature selection procedures require customized programs. However, as computation cost is getting cheaper each day, these problems are becoming less of an obstacle for modelers.

## REFERENCES

Ahn, B. –H. (1996). Forward additive neural network models. *Ph.D. dissertation*, Kent State University, Kent, Ohio, USA.

Belue, L. & Bauer, K. J. (1995). Determining input features for multilayer perceptrons. *Neurocomputing*, 7, 111-121.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford, UK: Oxford University Press.

Cottrell, M., Girard, B., Mangeas, M., & Muller, C. (1995). Neural modeling for time series: A statistical stepwise method for weight elimination. *IEEE Transactions on Neural Networks*, 6, 1355-1364.

Dasgupta, C. G., Dispensa, G. S., & Ghose, S. (1994). Comparing the predictive performance of a neural network model with some traditional market response models. *International Journal of Forecasting*, 10(2), 235-244.

Draper, N. & Smith, H. (1981). *Applied Regression Analysis*. New York: John Wiley & Sons.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition (2nd edition)*. San Diego, CA: Academic Press.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation,* 4, 1-58.

Hassibi, B. & Stork, D. (1993). Second order derivatives for network pruning: Optimal brain surgeon. In S. Hanson, J. Cown & C. Giles (Eds.), *Advances in Neural Information Processing Systems* (Vol. 5, pp. 164-171). San Mateo, CA: Morgan Kaufmann.

Hu, M., Hung, M. S., & Shanker, M. (2000). Estimating posterior probabilities of consumer situational choices with neural networks. *International Journal of Research in Marketing*, 16(4), 307-317.

Hu, M., Hung, M. S., Shanker, M., & Chen, H. (1996). Using neural networks to predict the performance of Sino-foreign joint ventures. *International Journal of Computational Intelligence and Organizations*, 1(3), 134-143.

Hu, M., Patuwo, E., Hung, M. S., & Shanker, M. (1999a). Neural network analysis of performance of sino-Hong Kong joint ventures. *Annals of Operations Research*, 87, 213-232.

Hu, M., Zhang, G., Jiang, C., & Patuwo, E. (1999b). A cross-validation analysis of neural network out-of-sample performance in exchange rate forecasting. *Decision Sciences*, 30(1), 197-216.

Hui, M. K. & Bateson, J. E. G. (1991, September). Perceived control and the effects of crowding and consumer choice on the service experience. *Journal of Consumer Research*, 18, 174-184.

Hung, M. S., Hu, M., Shanker, M., & Patuwo, E. (1996). Estimating posterior probabilities in classification problems with neural networks. *International Journal of Computational Intelligence and Organizations*, 1(1), 49-60.

Hung, M. S., Shanker, M., & Hu, M. Y. (2001). Estimating breast cancer risks using neural networks. *Journal of the Operational Research Society*, 52, 1-10.

Kumar, A., Rao, V. R., & Soni, H. (1995). An empirical comparison of neural network and logistic regression models. *Marketing Letters*, 6(4), 251-263.

Le Cun, Y., Denker, J., & Solla, S. (1990). Optimal brain damage. In D. Touretzky (Ed.), *Advances in Neural Information Processing Systems* (Vol. 2, pp. 598-605). San Mateo, CA: Morgan Kaufmann.

Lee, E., Hu, M. Y., & Toh, R. S. (2000). Are consumer survey results distorted? Systematic impact of behavioral frequency and duration on survey response errors. *Journal of Marketing Research*, 37(1), 125-134.

Lippmann, R. P. (1997, April). An introduction to computing with neural networks. *IEEE ASSP Magazine,* 4-22.

Lo, A. (1996). Recent advances in derivative securities: Neural networks and other nonparametric pricing models. *International Workshop on State of the Art in Risk Management and Investments*, NUS, Singapore.

Moody, J. & Joachim, U. (1992). Principled architecture selection for neural networks: Application to corporate bond rating prediction. In D. Touretzky (Ed.), *Advances in Neural Information Processing Systems* (Vol. 4, pp. 683-690). San Mateo, CA: Morgan Kaufmann.

Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). *Applied Linear Statistical Models*. Chicago, IL: Irwin.

Refenes, A. P. N., Abu-Mostafa, Y., Moody, J., & Weigend, A. (1996). *Neural Networks in Financial Engineering*. Singapore: World Scientific.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representation by error propagation. In D. E. Rumelhart & J. L. Williams (Eds), *Parallel Distributed Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.

*SAS User's Guide: Statistics* (1998). NC: SAS Institute.

Shanker, M., Hu, M., & Hung, M. S. (1996). Effect of data standardization on neural network training. *Omega*, 24(4), 385-397.

Shanker, M. S. (1996). Using neural networks to predict the onset of diabetes mellitus. *Journal of Chemical Information and Computer Sciences*, 36(1), 35-41.

Simonson, I. & Winer, R. S. (1992, June). The influence of purchase quantity and display format on consumer preference for variety. *Journal of Consumer Research*, 19, 133-138.

Tam, K. Y. & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, 38(7), 926-947.

Tang, Z. & Fishwick, P. A. (1993). Feedforward neural nets as models for time series forecasting. *INFORMS Journal on Computing*, 5(4), 374-385.

West, P. M., Brockett, P. L., & Golden, L. L. (1997). A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science*, 16(4), 370-391.

Wong, F. S. (1991). Time series forecasting using backpropagation neural networks. *Neurocomputing*, 2, 147-159.

Zhang, G., Patuwo, E., & Hu, M. (1998). Forecasting with artificial neural networks: The state of art. *International Journal of Forecasting*, 14(1), 35-62.