ELSEVIER

Intern. J. of Research in Marketing 16 (1999) 307–317

International Journal of
Research in
Marketing

www.elsevier.com/locate/ijresmar

Estimation of posterior probabilities of consumer situational choices with neural network classifiers

Michael Y. Hu [a,b,*], Murali Shanker [a], Ming S. Hung [c]

[a] *Kent State University, Kent, OH, USA*
[b] *Chinese University of Hong Kong, Hongkong, China*
[c] *Optimal Solutions Technologies, USA*

Received 8 June 1998; accepted 9 November 1999

Abstract

This study shows how neural networks can be used to estimate the posterior probabilities in a consumer choice situation. We provide the theoretical basis for its use and illustrate the entire neural network modeling procedure with a situational choice data set from AT&T. Our findings supported the appropriateness of this application and clearly illustrate the nonlinear modeling capability of neural networks. The posterior probability estimates clearly add to the usefulness of the technique for marketing research. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Neural networks; Classification; Consumer choices; Posterior probabilities

1. Introduction

In recent years, there has been an upsurge in the business applications of artificial neural networks (ANNs). These applications can be classified into two broad areas: classification and time series forecasting. Zhang et al. (1999) noted that over 50 academic papers have been devoted to the former, which deals with the task of assigning an object to one of known groups (as opposed to clustering where groups were unknown before analysis). Examples here include prediction of bank bankruptcies (Tam and Kiang, 1992), of success in joint ventures (Hu et al., 1996; Hu et al., 1999a), and of consumer choices (Kumar et al., 1995; West et al., 1997). Examples of forecasting include prediction of derivative/option, stock prices (Lo, 1996; Refenes et al., 1996), and forecasting of currency exchange rates (Hu et al., 1999b). Zhang et al. (1998) provided an extensive review of forecasting models using ANN.

A review of literature reveals that ANNs have not been fully accepted as part of the methodology tool box of market researchers. Only two applications (Kumar et al., 1995; West et al., 1997) can be identified in the leading marketing research journals. Both studies claimed superiority of ANN over logistic regression and discriminant analysis, measured in classification rates. The West et al. (1997) study is a more comprehensive approach. The ANN models were compared with traditional statistical models in

* Corresponding author. Department of Marketing, College of Business, Kent State University, Kent, OH 44242, USA. Tel.: +1-330-672-2750 ext. 326; fax: +1-330-672-2448; e-mail: mhu@bsa3.kent.edu

0167-8116/99/$ - see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0167-8116(99)00018-X

two data sets: one from a simulated choice situation and the other from published consumer patronage behavior. Each data set was decomposed into three separate parts: training (60%), validation (20%) and test (20%). Selection of the network architecture, in this case primarily the number of hidden nodes, was based on the results in the validation sample. The test sample was used to measure the predictive ability of the model. For variable selection, the authors suggested using logistic regression to come up with a 'super list' of variables with significant *t* statistics for inclusion in neural network models. Dasgupta et al. (1994) also relied on logistic regression for variable selection.

One frequent criticism of neural networks is that they cannot 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 variable and an 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 paper, such analysis is reported.

A situational consumer choice model was constructed to illustrate the various aspects of building neural networks to predict what product or service a consumer will choose. Like the two marketing studies cited above, our neural network is a classification model; but unlike them, our approach relies on the estimation of *posterior probability* instead of simply trying to classify a consumer's choice. The posterior probability, being a continuous variable, allows more interesting analysis of the relationships between consumer choice and the predictor variables.

In addition to using posterior probability for consumer choice modeling, we wish to bring to the attention of market researchers who may contemplate using neural networks two model building issues:

- *Model selection.* Selection of an appropriate model is a non-trivial 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 and 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 and equal or comparable predictive power. For neural networks, as mentioned above, parameter testings do not apply and therefore more computational intensive methods must be employed to determine the variables that should be included in a model. We offer a heuristic that seems to work well for the test data set.

The organization of this paper is as follows. In Section 2, a relatively extensive introduction of neural networks is given. Section 3 introduces the theoretical basis for the estimation of posterior probabilities. The entire approach to model situational choice prediction is illustrated in Section 4. The data came from a large scale study conducted by American Telephone and Telegraph (AT&T) in the early 1970s. As can be expected, one of the objectives of the study was to find out how consumers chose between various methods of communication, including long-distance telephone calls. The results are presented in Section 5 and the conclusions in Section 6.

2. Neural network models

ANNs are flexible, nonparametric models. A network is an abstract structure composed of nodes and arcs. Each arc connects two nodes, an origin and a destination, and information flows in the direction from the origin to the destination. The networks used in this study are called *feedforward* networks because the arcs do not form a circuit, in that as one travels from a node following the direction of the arcs, one cannot return to the starting node. A kind of feedforward network called *multi-layer perceptron* has nodes grouped into layers — the lowest layer is the *input layer* and the highest the *output layer*, and the layers in between are called the *hidden layers*. Arcs exist usually only between nodes of adjacent layers. In other words, nodes in a layer are only connected to the nodes in the next higher layer.

Each node gathers the signals coming through the arcs, then transfers the total input into an output. To be specific, let x_j and y_j stand for the input into and the output from node j, respectively. Then $y_j = a(x_j)$, where a is the transfer function (also called the *activation function*).

If node j is an input node, then x_j is given from outside the model and $y_j = x_j$. Otherwise,

$$x_j = \sum_{i \in S_j} w_{ij} y_i \tag{1}$$

where S_j denotes the (source) set of nodes connected into node j, w_{ij} is the weight on the arc from a source node i to node j. Some nodes may have an extra scalar (the intercept) on the right-hand-side of the equation. If node j contains a scalar, then we define S_j to include node 0 whose output is always 1 and use the weight w_{0j} as the scalar.

In our networks, the activation function in the hidden and the output nodes is the logistic function which has the form $a(x) = (1 + e^{-x})^{-1}$. Theoretically, a network with one hidden layer and logistic activation function at the hidden and output nodes is capable of approximating any function arbitrarily closely (Cybenko, 1989; Hornik, 1991, 1993). Hence, the networks used in this study all have one hidden layer and logistic activation in the hidden and output layers. However, we also allow for arcs from the input nodes to the output nodes. Fig. 1 shows an example network, where nodes 1 and 2 are input nodes; nodes 3 and 4 are hidden nodes, and node 5 is the output node. Arcs from node 0 represent the scalars to the destination nodes. In the example, it shows that the node 5 has a scalar.

The main reason for allowing direct arcs from input nodes to output nodes is as follows. Let x_j be the input to an output node j in a network with one hidden layer whose hidden nodes are represented by set H and whose input nodes are represented by set I. Then $S_j = \{0\} \cup I \cup H$ and

$$\begin{aligned} x_j &= w_{0j} + \sum_{i \in I} w_{ij} y_i + \sum_{k \in H} w_{kj} y_k \\ &= w_{0j} + \sum_{i \in I} w_{ij} y_i + \sum_{k \in H} w_{kj} a\left(\sum_{i \in I} w_{ik} y_i \right) \end{aligned} \tag{2}$$

where w_{0j} is the optional scalar. One can see that the direct arcs enable the inclusion of linear terms in the equation and hence allow for greater flexibility in the model.

The set of arc weights will be referred to as the parameters of the neural network model and will be denoted by $\boldsymbol{w} = (w_{ij})$. The parameters are determined by *training* the network with examples. Let $\boldsymbol{X} = (\boldsymbol{X}^p)$ and $\boldsymbol{T} = (\boldsymbol{T}^p)$ be the training data where $\boldsymbol{X}$ is a matrix of input variables, $\boldsymbol{T}$ a matrix of target values for the output nodes. $\boldsymbol{X}^p$ and $\boldsymbol{T}^p$ are a row of $\boldsymbol{X}$ and $\boldsymbol{T}$, respectively, where p stands for a pattern. The number of columns in $\boldsymbol{X}$ is equal to the number of input nodes of the neural network, and the number of columns in $\boldsymbol{T}$ is equal to that of the output nodes. For each p, $\boldsymbol{X}^p$ is given to the input nodes and after following the network structure, we will receive output $\boldsymbol{Y}^p$. (We show vectors and matrices in boldface.) Training is to select the parameters such that $\boldsymbol{Y}^p$ is as close to $\boldsymbol{T}^p$ as possible. The typical measure is the *sum of squared errors* (SSE):

$$\sum_p \sum_j \left(Y_j^p - T_j^p \right)^2 \tag{3}$$

where j denotes an output node.

Obviously we would like to minimize the SSE. Neural network training is thus a problem of nonlinear minimization — in fact, a least squares problem. Probably the most popular training method is the back-propagation (Rumelhart et al., 1986). In this study, we used the algorithm developed by Ahn (1996) which starts with a network with zero hidden nodes and assigns the parameter values found in linear regression as the starting values of the arc weights. Then a hidden node is added. The weights of the arcs in the previous network are used as starting weights. The starting weights for the new arcs are determined by a scheme designed to maximize the reduction in SSE. After that, a standard nonlinear optimization algorithm is used to find the

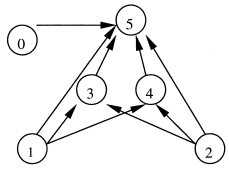

Fig. 1. Example neural network.

solution which minimizes SSE. Then another hidden node is added, until the desired (predetermined) number is reached.

3. Estimation of posterior probabilities

3.1. Definitions

A classification problem deals with assigning an object, based on its attributes, to one of several groups. Let $\boldsymbol{x}$ be the attribute vector of an object and ω_j denote the fact that the object is a member of group j. Then the probability $P(\omega_j|\boldsymbol{x})$ is called the posterior probability and it measures the probability that an object with attributes $\boldsymbol{x}$ 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 and 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.

In addition, with the posterior probability, a modeler can make assignments based on the minimum cost of mis-classification. Suppose a particular $\boldsymbol{x}$ is observed and is to be assigned to a group. Let $c_{ij}(\boldsymbol{x})$ be the cost of assigning $\boldsymbol{x}$ to group i when it actually belongs to group j. The expected cost of assigning $\boldsymbol{x}$ to group i is

$$C_i(\boldsymbol{x}) = \sum_{j=1}^{m} c_{ij}(\boldsymbol{x})\, P(\omega_j|\boldsymbol{x}). \tag{4}$$

The objective of a decision maker is to minimize the total expected cost and that is accomplished by

$$\text{Decide } \omega_k \text{ for } \boldsymbol{x} \text{ if } C_k(\boldsymbol{x}) = \min_i C_i(\boldsymbol{x}). \tag{5}$$

The above is known as the *Bayesian decision rule* in classification. In other words, assign object $\boldsymbol{x}$ to the group where the expected cost of mis-classification is the smallest.

A particular case for the rule is when the cost is binary: $c_{ij}(\boldsymbol{x}) = 0$ if $i = j$, and 1 otherwise. The cost function $C_i(\boldsymbol{x})$ can be simplified to

$$C_i(\boldsymbol{x}) = \sum_{i \neq j} P(\omega_j|\boldsymbol{x}) = 1 - P(\omega_i|\boldsymbol{x}), \tag{6}$$

and the Bayesian decision rule is reduced to

$$\text{Decide } \omega_k \text{ for } \boldsymbol{x} \text{ if } P(\omega_k|\boldsymbol{x}) = \max_i P(\omega_i|\boldsymbol{x}). \tag{7}$$

The general cost (4) is useful in applications where the cost of a mis-classification is different for different groups. For example, in the bank failure prediction model of Tam and Kiang (1992), for a depositor the failure to predict a bank going bankrupt could mean a greater cost than the mistake of declaring a healthy bank bankrupt. When the cost is equal or unavailable, the 0–1 cost can be used. Then, using rule (7), the decision is to minimize the number of mis-classifications.

3.2. Least squares estimator

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. Specifically, suppose the object of interest $\boldsymbol{x}$ is the pth object in the data set, then with objective function (3) and target values defined as follows,

$$T_j^p = \begin{cases} 1 & \text{if object } \boldsymbol{x} \text{ belongs to group } j \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

the least squares solution y_j^p has the property that $E(y_j^p) = P(\omega_j|\boldsymbol{x})$ for object $\boldsymbol{x}$; in other words, the predicted value for the jth output variable is an unbiased estimator of the posterior probability that $\boldsymbol{x}$ belongs to group j.

Neural networks provide a convenient way to perform this computation. For a classification problem with d features and m groups, the neural network structure will have d input nodes and m output nodes and the target values for the output nodes will be defined as in Eq. (8). However, for two group classification problems, only one output node is sufficient and the target values will be 1 for group 1 and 0 for group 2.

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.

3.3. 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.

4. Empirical examination of situational influences on choice

AT&T 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 3990 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.

4.1. 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 non-verbal, 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 copy. 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

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

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 be captured.

A sample of 1480 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.

4.2. Design of neural network models

As previously mentioned, the networks used in this study are feedforward networks with one hidden layer. Direct connections from the input layer to the output layer are added, for the reason explained earlier, right after Fig. 1. 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 is the scaling of input variables before training. Previous research (Shanker et al., 1996) indicates that data transformation is not very helpful for classification 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 and Fishwick, 1993), $2d$ (Wong, 1991), and $2d + 1$ (Lippmann, 1997), for a one-hidden-layer of d input nodes. However, none of these heuristics work 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 one hidden node and calculate the validation set 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.) Fig. 2 shows the plot of SSE for the validation set as the number of hidden nodes varies from 0 to 6, with all the 11 feature variables as inputs.

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

4.3. Selection of input variables

As discussed earlier, feature selection is an important and difficult topic in neural network modeling.

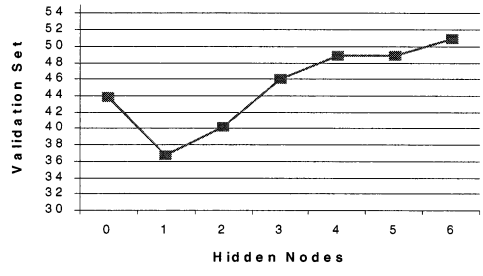

Fig. 2. Validation set SSE versus number of hidden nodes.

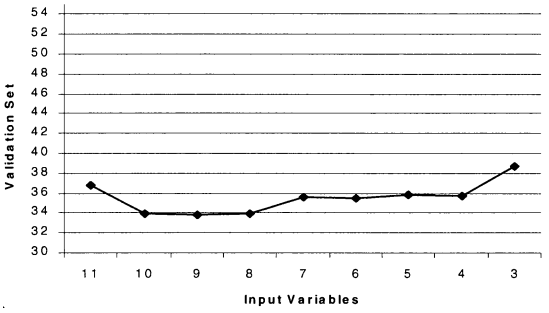

Fig. 3. SSE versus number of features.

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 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; Shanker, 1996; Hung et al., 1999).

As indicated in Fig. 3, the validation set SSE for the 11-variable model is around 37. It drops to around 34 for the 10-, 9- and 8-variable models. It increases to about 36 and remains there for 7- to 4-variable models. The next variable removal brings about a sharp increase in SSE. Although the 8-variable model has the smallest SSE, the 4-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 major impact on situation-based choices (Hui and Bateson, 1991; Simonson and Winer, 1992). Conceptually, one can expect situation-specific factors to exercise greater amount of 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.

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.

5. Results

5.1. 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 could carry out some preliminary analyses before the models were built. Two of the variables, REASON and RECEIVER, are 0–1 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 an 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. In the interest of brevity,

Table 2
Frequency table for COMMTYPE

RECEIVER	REASON		Total
	0	1	
0	59/83 = 0.711	191/545 = 0.350	250/628 = 0.398
1	84/103 = 0.816	240/453 = 0.530	324/556 = 0.583
Total	143/186 = 0.769	431/998 = 0.432	574/1184 = 0.485

only the scatter plot for COMMTYPE against TOTALCOM and NUMCALLS when REASON = 0 and RECEIVER = 0 is shown (Fig. 4). It is difficult to see any relationship between the dependent variable and either of the continuous variables.

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, the relationship patterns we failed to find in Fig. 4 are much clearer now, as shown in Fig. 5. First, the 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 3 plots and will not be presented here.

Table 3
Mean posterior probability

RECEIVER	REASON		Total
	0	1	
0	0.744	0.323	0.379
1	0.781	0.514	0.564
Total	0.764	0.410	0.466

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 over 30), then the 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.

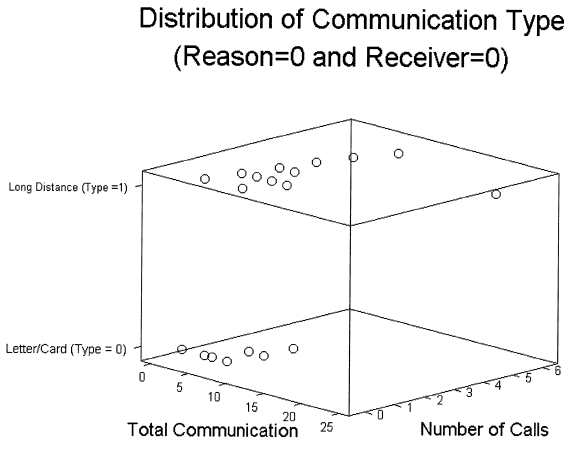

Fig. 4. Preliminary analysis.

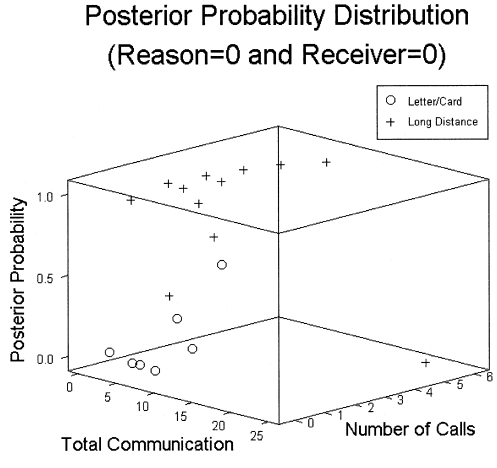

Fig. 5. Posterior probability function.

5.2. 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 eleven 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 eleven 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.

Table 4
Classification rates (correct classifications) for the test set

Model	Group	Neural network	Logistic regression
11 features	Total	0.818 (242)	0.787 (233)
	Long distance	0.870 (114)	0.817 (107)
	Letter/card	0.776 (128)	0.764 (126)
7 features	Total	0.818 (242)	0.804 (238)
	Long distance	0.763 (100)	0.817 (107)
	Letter/card	0.861 (142)	0.794 (131)
4 features	Total	0.831 (246)	0.794 (235)
	Long distance	0.840 (110)	0.779 (102)
	Letter/card	0.824 (136)	0.806 (133)

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.

6. 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 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 not on classification rates. Rather, it is on the estimation of these posterior probabilities and the 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 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. Dasgupta 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 an obstacle for modelers.

References

Ahn, B.-H., 1996. Forward additive neural network models. PhD Dissertation, Kent State University, Kent, OH, USA.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford Univ. Press, Oxford, UK.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematical Control Signals Systems 2, 303–314.

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.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural Computation 4, 1–58.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Networks 4 (2), 251–257.

Hornik, K., 1993. Some new results on neural network approximation. Neural Networks 6 (8), 1069–1072.

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. 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., 1999. Estimating breast cancer risks using neural networks. Working Paper, College of Business, Kent State University, Kent, OH, USA.

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

Lee, E., Hu, M.Y., Toh, R.S., 2000. Are consumer survey results distorted? Systematic impact of behavioral frequency and du-

ration on survey response errors. Journal of Marketing Research, in press.

Lippmann, R.P., 1997. An introduction to computing with neural networks. IEEE ASSP Magazine, April, 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.

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

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

SAS User's Guide: Statistics, 1998. SAS Institute, NC.

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. The influence of purchase quantity and display format on consumer preference for variety. Journal of Consumer Research 19 (June), 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., Hu, M., Patuwo, E., Indro, D., 1999. Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. European Journal of Operational Research 116 (1), 16–32.

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