Forecast of the Dynamics of the Undecided Population in a Public Opinion Poll by a Neural Network

C. Monterola*, M. Lim*, J. Garcia^, and C. Saloma*
*National Institute of Physics
University of the Philippines Diliman, 1101 Quezon City, Philippines
^Ateneo De Manila University, Quezon City, Philippines

EXTENDED ABSTRACT

One of the unsettled issues in any survey concerns the handling of the undecided respondents (UR). This problem can be safely ignored if UR constitute a small percentage of the population, and if their number is not enough to substantially sway the outcome of the survey. However, if the UR for example is about 30% and the decided respondents (DR) splitting 40% : 30%, then the results of the survey have to be taken as inconclusive.

The conventional way of dealing with this problem, as shown in print and television all over the Philippines, is simply to split the UR in proportion to the DR. Basically such undertaking merely 'erases' the possible significance of the UR. Another approach is to allocate the UR among the different responses, based on such criteria as geographical distributions (e.g., residents of a given area have traditionally gone with candidate X), or socioeconomic distributions (because, perhaps, affluent voters tend to go Republican in the U.S.)(Visser, 1996). Such may be termed an *exogenous* method, inasmuch as it relies on considerations other than the data at hand.

Still another method is to use statistical analysis tools, such as discriminant analysis (DA), which tries to look at patterns in statistical distributions to forecast the most probable separation of the undecided respondents. This has been used in the past, with claims of up to 86% success rate (Fenwick, 1982). Such a method we would term *endogenous* since it looks only at the given data and infer its conclusions only therefrom, with minimal or no heuristics based on outside considerations, beyond what may be found in the data itself.

In this study commissioned by Pulse Asia, Inc. (PA), a professional public opinion poll organization, we demonstrate that an artificial neural network (ANN) can forecast up to 94% accuracy the most apparent sentiment of the UR when he chose to decide, or when he is forced to decide. Specifically, an ANN is tasked to determine how the electorate rates the performance of Joseph Estrada, who was the incumbent president of the Philippines. The poll questionnaire consists of one direct question and 291 (pro-rated) indirect ones that probe into the opinions of the respondent on specific socioeconomic and political issues, and ratings of other government executives, legislators, and institutions. The answers to the direct question can be divided in three parts: one, those who approved of the Estrada administration, second those who disapproved, and third, those who are undecided. A total of 1200 respondents were chosen randomly all over the Philippines with a claim that the sampling procedure constitutes an uncertainty of ± 6% and ± 3% in the national and regional level, respectively (Pulse Asia, 2000)

ANN's are known to perform successfully in image recognition and feature identification in different kinds of signals. Compared with other statistical methods, the use of NN's in pattern recognition has the following advantages: (1) Recognition can proceed without any a priori knowledge about the characteristics of the classes into which the population is to be identified;

Science Diliman (July-December 2001) **13:2**, 104-107

(2) Recognition capability of trained NN's is robust against input noise; (3) Recognition by a trained NN is fast (Haykin, 1999; Soriano, 1995).

In recent years, ANN's have also been utilized to predict the behavior of complex economic systems found in stock trading and securities, financial management fraud, inflation, risk assessment in auditing, and financial earnings (Anders, 1988). In our literature search however, the use of NN's in forecasting the behavior of the undecided population in a political opinion poll has never been undertaken.

The ANN was used on data from two surveys run by PA, one in December 1999, and the other in March 2000. In the December 1999 survey of 1200 respondents: 596 approved of the Estrada administration (P); 332 disapproved of the Estrada administration (N); and 272 were undecided (U). Of the 928 DR, we selected 700 at random, and used this set to train the ANN. The training was for recognizing whatever pattern(s) there may be in the respondents' replies to the twenty peripheral questions which serve as the inputs $\{x_1, x_2, \ldots x_N\}$ of the ANN, and their relation to the replies to the main question which serves as the output y_{NN} of the ANN. The peripheral questions involve the approval ratings of other government executives, legislators, and institutions. Inquiries regarding the socio-demographic conditions (age, educational attainment, civil status, etc.) of respondent form a separate part of the questionnaire. The answers (r) to the indirect questions related to approval ratings are pro-rated numerically in the following manner: r=5 (strongly supportive); r=4 (mildly supportive); r=3 (no opinion), r=2 (mildly unsupportive), r=1 (strongly unsupportive). For questions dealing with socio-demographic data, the corresponding numerical ratings to r are based on the actual number of possible choices of the respondents.

Using a feedforward supervised backpropagation learning method (SBLM) the ANN is trained such that it minimizes $E = |\psi^{(k)}_{correct} - \psi^{(k)}_{NN}|^2$, where $\psi^{(k)}_{correct}$ is the correct response at the kth data set and $\psi^{(k)}_{NN}$ is the approximate response of the network given by $\psi_{NN}^{(k)} = f_0(S_{m=1,H}\, d_{sm}^{(k)} y_m^{(k)})$; $y_m^{(k)} = f_0(S_{j=1,L}\, w_{mj}^{(k)} x_j)$ after the presentation of the kth data set (Haykin, 1999). The m^{th} output of the hidden nodes is $y_m^{(k)}$ $\{m=1,2 \ldots H\}$ and the hidden and output activation functions are represented by $f_H(z)$ and $f_0(z)$, respectively. The free parameter $w_{mj}(k)$ describes the interconnection weights between neurons in the m^{th} and the j^{th} layer, while $d_{sm}(k)$ gives the synaptic strength of the s^{th} and m^{th} layer. Minimization of E using SBLM proceeds by updating the weights w_{mj} according to $w_{mj}^{(k+1)} = w_{mj}^{(k)} - \eta \partial E^{(k)}/\partial w_{mj}^{(k)} - \alpha\, \partial E^{(k-1)}/\partial w_{mj}^{(k-1)}$. Where η and α are known as the learning and momentum term that are adaptively varied both to hasten the reduction of E and to prevent traps at local minimum (Haykin, 1999). Similar update rules hold for d_{sm}.

Our results have shown that the optimum architecture is that with three layers, 30 hidden nodes (H=30) and $f_H(z) = f_0(z) = 1.7159 \tanh(2z/3)$. After training, the remaining 228 responses, which make up the *test set*, were then fed into the ANN which, using the patterns to the replies that it has "learned", was expected to predict each response to the main question . The ANN predicted correctly the responses to the main question 93.86% of the time, or 214 out of the 228 respondents in the test set.

For the March 2000 survey again with 1200 respondents, the ratio of P:N:U is 559 : 357 : 284. Exploiting a similar procedure outlined above 700 DR were chosen at random for training the ANN, leaving 216 to comprise the test set. In this instance, the ANN was able to predict correctly 93.52% of the responses to the main question in the test set (or 201 out of 216 respondents) after it has been trained. But the success rate was raised even higher to 95.37% (or 206 out of 216 respondents) when responses to four of the peripheral questions were replaced by four demographic responses. By computing the success rate fluctuations of 50 different random choices of DR, we verify that this improvement is not merely a result of statistical variability. Hence, this indicates that there is relevant information in the demographic questions that are not found in the peripheral questions.

Next, we confirm that the UR belong to the same population as the decided ones insofar as their replies to the peripheral questions are concerned. It must be stressed that the ANN was trained using the replies to the peripheral questions of the decided respondents. Hence, if the undecided respondents make up a different population from the decided ones - with respect to the peripheral questions - then using the ANN to predict the sentiments of the undecided respondents might not be valid.

Consequently we looked at the clustering of all the respondents in a 20-dimensional space which represent the possible replies to the 20 peripheral questions. We found that the undecided respondents do not form a cluster by themselves, distinct from that of the decided respondents.

The distribution of the survey data with respect to the 20 peripheral questions is analyzed by constructing a 20-dimensional ellipsoids centered at V_P (or V_N) where the axes are given in units of the standard deviations σ_{Pi} (or σ_{Ni}), i = 1, 2, ..., 20. The coordinate V_{Pi} (or V_{Ni}) is the mean value of the response of the approved (**P**) (or disapproved **N**) class to the i^{th} peripheral question. Fig. 1 plots the percentage of respondents that are contained in the ellipsoid with axes given in units of σ_{Pi} (or σ_{Ni}). As an example, 25% of the approved (**P**) respondents and 25% of the undecided (**U**) respondents

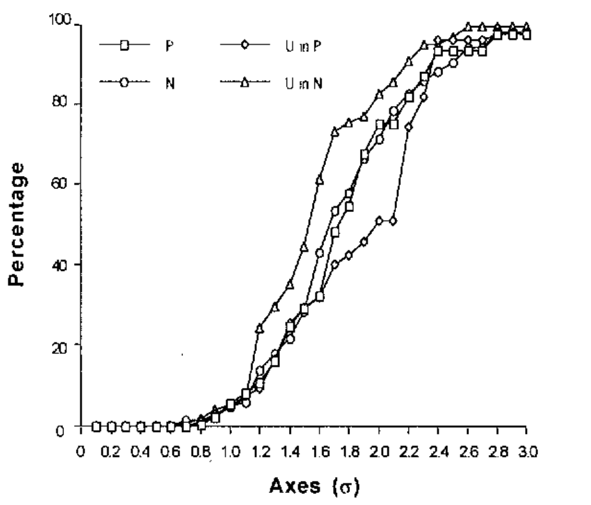

Fig. 1. Percentage profile of the distribution of: (a) P respondents within the V_P-centered ellipsoids (square), (b) U respondents within the V_P-centered ellipsoids (diamond), (c) N respondents within the V_N-centered ellipsoids (circle), (d) U respondents within the V_N-centered ellipsoids (triangle)

fall within the 1.4σ - ellipsoid centered at V_P, i.e., it is in the range $V_P \pm 1.4\ \sigma_{Pi}$. Fig. 1 shows that 97% of the U respondents fall within the 3σ-range of the V_P-centered ellipsoid, while 99% of the same U respondents fall within the the 3σ-range of the V_N-centered ellipsoid. Hence, insofar as the peripheral questions are concerned, the decided and the undecided respondents come from the same population, making the use of ANN valid.

So we next turned to the undecided respondents, feeding the ANN their responses to the peripheral questions. The ANN made the following predictions: For the December 1999 survey, the ANN split the 272 undecided respondents into: 172 ± 7 would be P and 100 ± 4 would be N or a ratio of about 1.72:1. Note that for the DR, the ratio is about 1.80:1.

For the March 2000 survey, the ANN split the 284 undecided respondents into: 175 ± 3 would be P, 109 ± 2 would be N or a ratio of about 1.61:1. Note that for the DR, the ratio is about 1.57:1. In both cases, interestingly, the ANN predicted that the undecided respondents, if they harbor any sentiment at all, would do so in just about the same approval/disapproval ratio as the decided respondents.

Finally, we compare our results by utilizing a multivariate numerical method known as discriminant analysis to separate the UR into either the P or N states. The resulting accuracy for both the December 1999 and the March 2000 surveys are about 70%. Moreover, it predicts that the undecided respondents in the December survey would split as follows: 136 would be P and 136 would be N. For the March survey, the discriminant analysis DA gives the following ratio for P:N is 146:128. The results are nowhere close to those obtained by the ANN.

REFERENCES

Anders, U., O. Korn, & C. Schmitt, 1998. Improving the pricing of options: a neural network approach. *J. of Forecasting*. 17:369-388.

Charitou, A. & C. Charalambous, 1996. The prediction of earnings using financial statement information: empirical evidence with logit models and artificial neural networks. *Int. J. of Intelligent Systems in Accounting, Finance & Management.* 5:199–215.

Fanning, K. & K. Cogger, 1998. Neural network detection of management fraud using published financial data. *Int. J. of Intelligent Systems in Accounting, Finance & Management.* 7:21-41.

Fenwick, I., F. Wiseman, J.F. Becker, & J.R. Heiman, 1982. Classifying undecided voters in pre-election polls. *Public Opinion Quarterly.* 46: 383-391.

Garcia, R. & R.Gencay, 2000. Pricing and hedging derivative securities with neural networks and a homogeneity hint. *J. of Econometrics.* 94:93.

Gençay, R . & T. Stengos, 1998. Moving average rules, volume and the predictability of security returns with feedforward networks. *J. of Forecasting.* 17:401-414.

Hamm, L. & B. Brorsen, 2000. Trading futures markets based on signals from a neural network. *Appl. Econ. Lett.* 7:137.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. 2nd Ed. New York, Prentice-Hall.

Moody, J., L.Wu, & Y. Liao, 1998. Performance functions and reinforcement learning for trading systems and portfolios. *J. of Forecasting.* 17:441-470.

Pulse Asia, Inc., 2000. Ulat ng Bayan Survey. Philippine Social Science Center Building, Commonwealth Ave., Diliman, Quezon City, Philippines.

Soriano, M. & C. Saloma, 1995. Cell classification by a learning principal component analyzer and a backpropagation neural network. *Bioimaging.* 3: 168.

Visser, P., J. Krosnick, J. Marquette, & M. Curtin, 1996. Mail surveys for election forecasting? An evaluation of the Columbus dispatch poll. *Public Opinion Quarterly.* 60:181-227.