Engineering Letters, 13:1, EL_13_1_3 (Advance online publication: 4 May 2006)

A framework for neural network to make business forecasting with hybrid VAR and GA components

S. I. Ao, *Member, IAENG*

***Abstract*— Applying Vector Autoregression (VAR) and genetic algorithm (GA) in hybrid systems with neural network can improve the NN's prediction capability. Two case studies have been carried out to demonstrate how to build our VAR-NN-GA system and its advantages. One is on the tourist patterns. Relatively recently, neural network is introduced into the tourist forecasting field [1, 2, 3]. Results show that the hybrid forecasting system is more robust and able to select variables automatically and makes more accurate prediction than the stand-alone neural network. Another case study is the Asian Pacific stock markets, which is more complicated. There exist strong time-dependent correlations between the US market and the Asian markets. While it is difficult to model the markets' interactions with a static model, VAR and GA are adopted for the dynamic model-selection process. The results for the stock market case study show that the hybrid system is about 30% better than the best individual NN for this case study. We can see that, as the domain of prediction become more complicated, the advantage of our system is clearer.**

***Index Terms*— Vector Autoregression (VAR), genetic algorithm (GA), hybrid systems, neural network, business forecasting.**

I. INTRODUCTION

The hybrid Vector Autoregression, neural network and genetic algorithm (VAR-NN-GA) framework can supplement its separate stand-alone components. The aim of the system here is to automate the decision process of prediction. The framework of the VAR-NN-GA is as followed (fig. 1):

(1) VAR analysis, which is to search for the correlated and leading indicators automatically;

(2) Neural network prediction, which is to make forecasting from the relevant inputs decided by the VAR analysis;

(3) Evolutionary process, which is to cope with the time-dependent nature of the co-relationships among the variables and to adjust the weightings of each neural network model.

Generally speaking, this VAR-NN-GA hybrid system can work for inputs data like indices of the regional countries and other business indicators. Input data, such as the trading volume, economic growth rate and currency exchange rate, etc., can also be tested in the VAR analysis. For the input variables with the lowest significance level, they will be used as the input variables for the neural network. In many cases, the performance of the predictions made by the neural network with these input variables is time-dependent and unstable. In some sub-periods, some input variables may be fitter for the prediction, while, in other sub-periods, they may be poorer in fact.

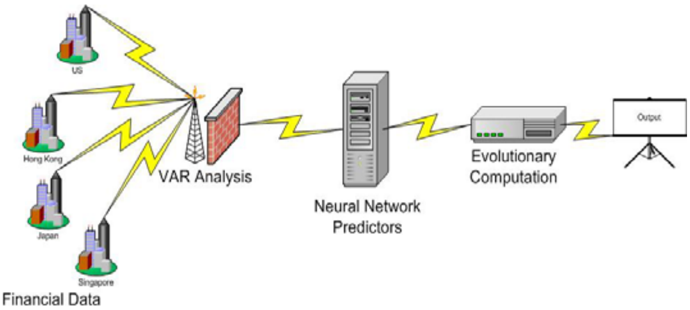

Fig. 1. VAR-NN-GA System Diagram

We can regard these neural network prediction models with different inputs as experts of different opinions of the relevant input factors. Their predictions may vary with time, as said. The selection and evaluation of these predictors can be made in the evolutionary cycle. Experts with higher forecasting accuracy in each cycle are going to weigh more heavily in the coming round. The detailed algorithm of our VAR-NN-GA system is as followed:

Algorithm 1. *VAR-NN-GA system*

```
Input: Multivariable Time Series Data (MTS) like index, visitor number, GNP, etc.
1. Pass the MTS data through VAR test;
2. Test the variables against each other to see their respective significance levels;
3. While the lag terms of variables are within the confidence interval,
        Select these terms;
```

Manuscript received December 8, 2005.

S. I. Ao can be contacted with e-mail: ao@alumni.cuhk.net

4. Form the input vectors for the neural network from the above selected MTS;
5. Make predictions by neural network from these different selected MTS;
6. Select the fittest prediction model for each sub-period by GA.

II. METHODOLOGY

2.1 Vector Autoregression

The Vector Autoregression (VAR) techniques [4, 5] are used to assist the understanding of their interactions with each other. The basic idea of VAR is to treat all variables symmetrically. Simply, VAR is a multivariate system of equations that we do not need to take the dependence versus independence into account.

In a VAR model of n variables, let ε_{it} denote the independent disturbances, C_i be constants y_{it} , $i = 1,...,n$ be the n variables at time t . The following equation shows us the methodology of VAR:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{pmatrix} = \begin{pmatrix} A_{11}(L) & \cdots & A_{1n}(L) \\ \vdots & \ddots & \vdots \\ A_{n1}(L) & \cdots & A_{nn}(L) \end{pmatrix} \begin{pmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{pmatrix} + \begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{nt} \end{pmatrix} \tag{1}$$

where $A_{ij}(L)$ takes the form of $\sum_{k=1}^{m} a_{ijk} L^k$, where L is the lag operator defined by $L^k y_t = y_{t-k}$, and m is the lag length specified by the user.

VAR's advantage is that multiple variables can be investigated at the same time. This characteristic is suitable for our research to study the interactions among the Asian Pacific markets and the tourist markets without pre-defined assumption.

2.2 Neural Network

The idea for the neural network is to mimic the working of our brain. It consists of axons for inputs, synapses, soma, and axons for outputs [6, 7]. In the typical neural network, there are three layers-the input layer, the hidden layer and the output layer. All these layers are connected and the architecture of the neural network design is itself a worthy field. To simplify the study and to make comparison easier with other studies, the model employed is the back-propagation method as followed:

Algorithm 2. Back-propagation algorithm of NN

1. Present the input vector patterns to the network;
2. Propagate the signals forwards, and calculate

$$u_j = a_{0j} + \sum_{i=1}^{I} a_{ij} x_i, \quad v_k = b_{0k} + \sum_{j=1}^{J} b_{jk} y_j,$$

$$y_j = g(u_j), \quad j = 1,..., J, \quad z_k = g(v_k), \quad k = 1,..., K$$

3. Calculate the mean squared error

$$E = \frac{\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(z_{kn} - t_{kn}\right)^2}{NK}$$

4. Update the weights according to the delta rule:

$$w^{m+1} = wm - \lambda d^m, \quad d^m = \sum_{n=1}^{N}\left(\frac{\delta y}{\delta x}|_m\right)_n$$

5. Repeat the above steps 2, 3, 4 until the error is less than the predefined value or for a predefined number of iterations.

The advantage of the neural network is that we do not necessarily predetermine the relationship between inputs and outputs with the exact functional form. Instead, it is decided by the data. Theoretically, it can approximate any functional forms of the input-output pair and can be used in the regression analysis.

There are criticisms that, as the relationship among the variables is not known in advance, the network acts just like a black box. This is one motivation for the proposed hybrid system of neural network and econometrics and genetic algorithm. With the VAR component, we can be assured that all of our feeding variables for the network have significant influences on the output variables. The GA component can enable us to tell others that our model will always select the fittest models among the different models.

2.3 Genetic algorithm

The idea of genetic algorithm (GA) is inspired by the concept of natural evolution, which is formulated by Charles Darwin in the 19th century. GA can be regarded as a broad collection of stochastic optimization algorithms that let the fittest to survive and the weak to die [8, 9].

In GA, the whole solution sets are called the population while an individual solution is referred to as a chromosome. In a chromosome, there exist different characteristics that are represented as the gene. They are corresponding to the different properties of an individual. There exist many generations in GA. The individuals will try to reproduce in each generation. The survival of the genes will follow the process of what has said in the above section. The procedure codes of GA are as followed:

Algorithm 3. *GA component*

1. Initialize P(t);
2. Evaluate P(t);

3. Recombine P(t) to yield C(t);
4. Evaluate C(t);
5. Select P(t+1) from P(t) and C(t);
6. Repeat the above three steps (3, 4, 5) in the next generation t+1 until the termination condition is met.

where t is the order of generation, P(t) is the population set at the generation t, C(t) is the population set after reproduction in the generation t.

III. CASE STUDY I: TOURIST INDUSTRY

The relationships between the current visitor number and its previous numbers (the time series analysis) are going to be studied. And the VAR is used to assist us with identify lead-lag dynamics among the variables.

3.1 Correlation Analysis of the Tourist Time Series Data with Macroeconomic Time Series Data

The visitor number time series data of various source countries are tested with its own lagged terms as well as against the time series of population and GDP. The hypothesis is that the variables are independent. And the significance level tells us the probability that the hypothesis is found invalid. For example, with significance level at 0.022 for the variable HKVISJAPF{2} on HKVISJAPF, it means that there is only statistically 2.8% probability of the independence of these two variables. In the order words, they are likely to be correlated.

From the identification process by the VAR tests, it is found out that the first lag term is the most suitable one for prediction by NN. And, the first lag term of the population size of the source market is also statistically significant with the visitor number model. This is different from the trial-and-error approach of training and predicting with the NN.

Let's look deeper into a tourist market, here Hong Kong as an example. During our modeling period, we can see that relationships among the significant variables are still far from static and it can be observed that the correlation is time-dependent. The following figure shows us the correlation among US visitors and the US population. At the time of low correlation, visitor number is mainly independent of the US population, while, during other periods, the two factors are highly correlated.

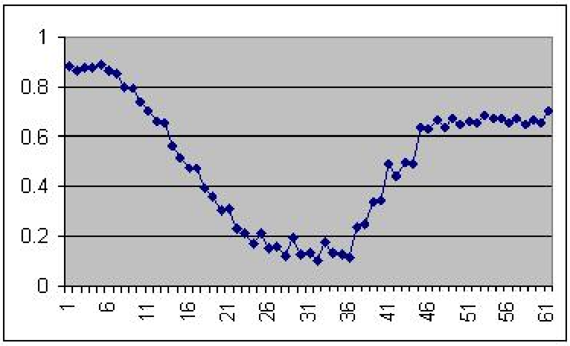

Fig. 2. Correlation of US visitor number with US population, using the correlation formulas of 40 sample data

3.2 Representing the Experts as Chromosomes of GA

As seen above, the suitability of feeding variables for network may change over time. We are going to represent the corresponding weightings of the input variables by GA. And their respective weightings at each sub-period are decided by number of its genes in the chromosome. For example, the below figure shows us an an expert of inputs of its own lag 1 term, and, another of inputs lag 1 term plus the population lag 1 term. They are represented as gene type 1 and gene type 2. In parent chromosome 1, the ratio is as 1:1. In parent 2, the ratio of the opinion of importance is 3:5. The child 1 is reproduced from the crossover of parent 1 and parent 2, by inheriting the first half of genes from parent 1 and the second half from parent 2. The child 2 is formed from the mutation process, by stochastically selecting the third and forth genes of parent 1 to change to other values. This mutation is to ensure that the population will be able to cover all possible opinions and has a globally suitable solution.

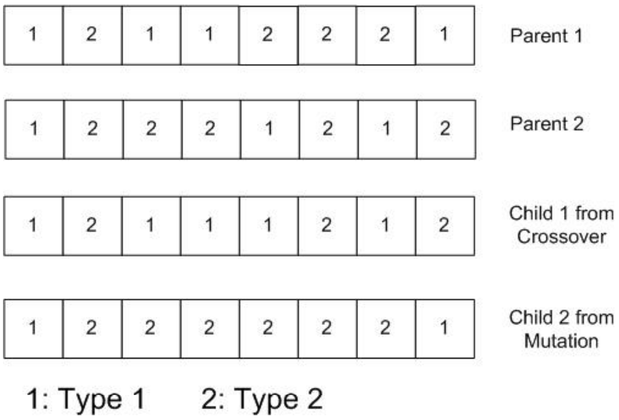

Fig. 3. Example of how to form the chromosome with genes, representing the two different models

3.3 EXPERIMENTAL RESULTS: A DEMOSTRATION OF THE ADVANTAGES OF THE SYSTEM

The monthly and quarterly tourist data for visiting Hong Kong have been used. The data is from January 1978 to December 2002, available from the DataStream. Separate origins of the visitors are available for those from Japan, South East Asia, West Europe, USA, Australia and New Zealand, Canada, and others, donating by HKVISJAF, HKVISASIF, HKVISWEF, HKVISUSA, HKVISANZF and HKVISCANF respectively. The total arrival number is donated by HKARRIVL. The GDP and population data is from the International Financial Statistics.

Results of the predictions made by neural network are as followed in the below table. It consists of two periods 1 and 2 respectively. In the period 1, the first 19th to 49th data are used for training the neural network and the next 10 are for testing. In the period 2, the first 59th to 89th data are used for training the neural network and the next 10 are for testing. It can be observed no unique model can outperform others over all the sub-periods. At the time of low correlation among visitor number and population (period 1), predictions with inputs of both US visitor and population (US_POP) under-performs. While at the period 2, the situation is reverse.

Table 1 Prediction of the US visitor number by NN for different time periods

Prediction	US_POP_1	US_1	US_POP_2	US_2
Mean Abs. Error	30569	25338	25522	33828
Mean Visitor Number	163990	163990	244288	244288
Percentage Error	18.64%	15.45%	10.45%	13.85%

The problem is that we can’t obtain the current correlation data before the prediction is made. It is not possible to automate the process to decide which model will be employed. Human expert’s opinion is needed to make the decision. Here, the genetic algorithm is employed to simulate this process.

From the following table of the tourist prediction, our hybrid NN-GA is better than the stand-alone neural network models. The data covers both periods 1 and 2. Results of label 1 and 2 are the sub-period results respectively. When the domains concerned become more complicated, for example, the stock markets, the advantage of the system will become clearer as shown in the following section.

Table 2 Prediction of the US visitor number by NN of inputs US visitor number & population, and, of US visitor number by hybrid NN-GA for the combined time period

Prediction	US_POP	US	NN-GA
Mean Abs. Error 1	30569	25338	25560
Percentage Error 1	18.64%	15.45	15.59%
Mean Abs. Error 2	25522	33828	29419
Percentage Error 2	10.45%	13.85%	12.04%
Mean Abs. Error	56091	29583	27489
Percentage Error	13.74%	14.49%	13.47%

IV. CASE STUDY II: ASIAN PACIFIC STOCK MARKETS

4.1 Analysis of the Lead-Lag Dynamics among the Markets

The stock data cover the stock markets in US and East Asia region, namely HK, JP, AU, SG, NASA100, PSCOMP and DJINDUS [10,11,12,13]. The data is in daily format from 3rd May 1990 to 3rd May 2002, available from the DataStream and Reuters. Our VAR results of the correlations among the Asian Pacific markets are summarized as followed (their dependence is tested with 5% significance level):

(1) HK depends on its past price, JP, Nasdaq, S&P and DJ;
(2) AU depends on its past price, S&P and DJ;
(3) SG depends on its past price, HK, Nasdaq, S&P and DJ;
(4) JP depends on its past price, Nasdaq, S&P and DJ;
(5) Nasdaq depends on its past price only;
(6) DJ depends on its past price and Nasdaq;
(7) S&P depends on its past price and Nasdaq;

Using the above VAR results, we can know which variables are most suitable inputs for the neural network. For the Asian markets, the relevant information is its own historical value as well as the stock movements from the US markets. But, such relationship is far from static and it can be observed that the correlation is time-dependent. For example, the following figure shows the changes of Hong Kong’s correlation with S&P over the recent ten years. Further investigation can tell us that, at the time of low correlation like the late 90’s of the Asian Financial crisis, the Hong Kong market (and similarly other Asian markets) is dominated by the local events like the currency problems. At other periods, the local markets are greatly correlated with the US markets.

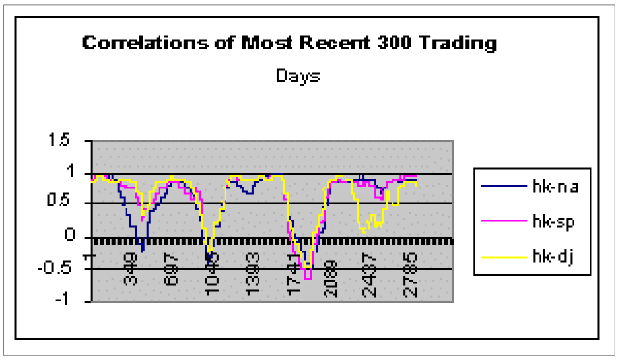

Fig. 4. Correlation of Hong Kong index with US's S&P, using the correlation formulas of 300-trading days

4.2 Results of Employing the Neural Network Prediction Models

Results of the predictions made by neural network are as shown in the following table. It consists of two periods 1 and 2 respectively. It can be observed that at different periods, the prediction made with inputs of both HK and S&P can outperform that with input of HK data only at the period 1. Period 1 is the time with high correlation between these two markets. At the time of low correlation like period 2, the situation is reverse.

The problem is that we can't obtain the current correlation data before the prediction is made. It is not possible to automate the process to decide which model will be employed. Human expert's opinion is needed to make the decision. Here, the genetic algorithm is employed to simulate this process.

Table 3 Prediction of the Hong Kong stock market by NN for different time periods

PREDICTION	HK_SP_1	HK_1	HK_SP_2	HK_2
Mean Abs. Error	46.24	96.11	737.47	314.59
Mean Index Value	3656.96	3656.96	14904.71	14904.71
Percentage Error	1.26%	2.63%	4.95%	2.11%

4.3 Comparison of Results from the VAR-NN-GA System with Stand-alone Neural Network

The below figure shows us a set of artificial sample data and demonstrates how the GA works in our system. There are two sets of predictions, let them be Expert 1 and Expert 2. They have different prediction values which are different from the actual value t[n]. Expert GA (EC) is the prediction made by the GA. We can see that Expert GA is closer to the actual values over the whole period. It can be observed from the following table, that the result made by GA, which is a combination of the Experts 1 & 2, works better than its individual components.

Table 4 Prediction of the Sample Data by Experts 1, 2 & GA respectively

PREDICTION	Expert 1	Expert 2	Expert GA
Mean Abs. Error	22.98	62.30	19.37
Mean Index Value	2968.31	2968.31	2968.31
Percentage Error	0.77%	2.10%	0.65%

For the real-life stock market prediction testing below, the model with hybrid NN-GA is better than the stand-alone neural network model. The data here covers both periods 1 and 2 of the table 1. Results of label 1 and 2 are the sub-period results respectively.

Table 5 Prediction of the Hong Kong stock market by NN of inputs HK & SP, HK and by hybrid NN-GA for the combined time period

PREDICTION	HK_SP	HK	NN-GA
Mean Abs. Error 1	46.24	96.11	42.03
Percentage Error 1	1.26%	2.63%	1.15%
Mean Abs. Error 2	737.47	314.59	286.83
Percentage Error 2	4.95%	2.11%	1.92%
Mean Abs. Error	391.85	205.39	164.43
Percentage Error	3.11%	2.37%	1.54%

V. CONCLUSION

Our hybrid VAR-NN-GA system for the prediction is designed to automate the process of selecting input variables, predictions and the evaluations of various prediction models as a whole. By varying the weightings of the input variables, we can have different prediction models. It is desirable to employ the most suitable model for each sub-period. This task is achieved by the genetic algorithm.

For the tourist demand forecasting, different models, mainly the econometrics and the neural network, have been developed separately for making prediction with this information. There are the linear systems like many econometric techniques. And there is also the nonlinear system like neural network, which has been known for its capability of pattern recognition and has played a more and more active role in the forecasting field. In the stock market analysis, previous studies have focused more or less on the historical prices and the trading volume of one market only. Our VAR-NN-GA system has the advantage of both the econometrics, which can offer clear explanation and testing for the correlations among the variables, and the GA, which can make our system become adaptive to the changing environment.

References

[1] Vincent Cho, *A comparison of three different approaches to tourist arrival forecasting*, Tourism Management, 24 (2003).

[2] Rob Law, Norman Au, *A neural network model to forecast Japanese demand for travel to Hong Kong*, Tourism Management 20 (1999), 89-97.

[3] Johann Du Preeez, Stephen F. Witt, *Univariate versus multivariate time series forecasting: an application to international tourism demand*, International Journal of Forecasting 1 (2002), article in press.

[4] Walter Enders, *Applied Econometric Time Series*, Wiley, 1995.

[5] William H. Greene, *Econometric Analysis*, Prentice Hall, 2000.

[6] Nikola K. Kasabov, *Foundations of neural networks, fuzzy systems, and knowledge engineering*, The MIT Press, 1997.

[7] A. S. Lapedes and R. Farber, *Non-linear Signal Processing Using Neural Networks: Prediction and System Modeling*, In Technical Report LA-UR-87. Los Alamos National Laboratory.

[8] A. Engelbrecht, *Computational Intelligent - An Introduction*, Wiley, 2002.

[9] A. Tettamanzi, M. Tomassini, *Soft Computing, Integrating Evolutionary, Neural, and Fuzzy Systems*, Springer, 2001.

[10] Ling T. He, *Time variation paths of international transmission of stock volatility-US vs Hong Kong and South Korea*, Global Finance Journal, 12, 2001, 79-93.

[11] J. W. Baek, S. Z. C, *An up-trend detection using an auto-associative neural network: KOSPI 200 futures*, Proc. Intelligent Data Engineering and Automated Learning 2002, Hong Kong.

[12] Ajay Samant and Korth. *American Depositary Receipts: The Performance of ADRs from the Asia-Pacific Region*, Journal of Asia-Pacific Business, 2(3), 1998.

[13] B. R. Oscar, S. R. Simon, F. R. Fernando, *Non-Linear forecasting methods: some applications to the analysis of financial series,* Progress in Economics Research II (Nova Science, 2002, P77-96).