GENETIC ALGORITHM FORECASTING FOR TELECOMMUNICATIONS PRODUCTS

STEPHEN D. SLOAN, RAYMOND W. SAW, JAMES J. SLUSS, JR., MONTE P. TULL, AND JOSEPH P. HAVLICEK
School of Electrical & Computer Engineering
The University of Oklahoma, Norman, Oklahoma

ABSTRACT
In this paper, we describe genetic algorithms (GA's) for forecasting long-term quarterly sales of products in the telecommunications technology sector using widely available economic indicators such as Disposable Personal Income and New Housing Starts as independent variables. Individual chromosomes indicated inclusion or disinclusion of specific economic variables, as well as operational rules for combining the variables. Population evolution utilized random crossover mating, mutation, and inversion. Several features beyond those of the canonical GA were also incorporated, including evolution of individuals in distinct ecosystems with a specified level of intermarriage between ecosystems, the capability for a single gene in an individual's chromosome to indicate a subroutine call to the complete chromosome of an individual from a previous generation, and hill-climbing applied to improve the most fit offspring produced by a generation. At a forecast interval of eight quarters, individuals exhibiting maximal fitness achieved RMS forecast errors below the the average two-week sales figure.

1. INTRODUCTION

The ability to accurately forecast long-term future sales of specific products is a highly desirable capability for many companies operating in the increasingly volatile telecommunications technology sector. Such a capability could allow companies to avoid surpluses and shortages in manufacturing resources, including materials, capital equipment, and personnel. Here, by *long-term*, we mean forecasting of quarterly sales at a forecasting interval ranging from a few quarters up to a few years. Recently, such long-term forecasting has proven to be an extremely difficult problem due to increased market volatility brought on by numerous factors including deregulation, the Telecommunications Act of 1996, and ever expanding global competition. Whereas simple heuristic location estimation techniques, including, *e.g.*, exponential smoothing, were in the past at least marginally adequate for developing long-term predictions in this market, we have found them to be wholly inadequate in recent years. Moreover, due to the highly nonstationary, evolutionary, and indeed sometimes even chaotic nature of telecommunications product sales, the effort required to continuously reformulate more sophisticated parametric methods including linear regressions, classical Box-Jenkins ARMA models, and Kalman filters can rapidly constitute an insurmountable burden.

Recently, artificial neural networks (ANN's) have been applied to a variegated array of forecasting problems (Donaldson and Kamstra, 1989), (Saravanan, 1993), (Kuan and White, 1994), (Wan, 1994), (Masters, 1995). One of the primary advantages of ANN's is that they potentially have the capability to capture complex and nonlinear relationships between the independent and forecasted variables, whereas such relationships are often difficult or impossible to treat using more traditional linear methods. The most important disadvantage of forecasting with ANN's is that it is generally impossible to formulate a deterministic description of the algorithm that is implemented by the network after training. Particularly in applications involving financial time series, this can be a significant concern since managers are often reticent to stake millions of U.S. dollars on a forecast that they cannot explain or systematically justify.

A related family of empirical optimization techniques that also have been applied recently to time series forecasting includes genetic algorithms (GA's) and evolutionary programming (Holland, 1975), (Packard, 1990), (Koza, 1992), (Bäck, *et al.*, 1997), (Mitchell, 1998). For example, Meyer and Packard (1992) proposed methods for prediction of high-dimensional chaotic time series using genetic algorithms based on conditional intervals for the independent variables. With this approach it is difficult to formulate a deterministic model relating the forecast and independent variables, however. Hybrid techniques have also been investigated, including both genetic-fuzzy systems (Goonatilake, *et al.*, 1994), (Kim and Kim, 1997) and the use of GA's in training ANN's (Saravanan, 1993), (Harrald and Kamstra, 1997). Some of these evolutionary techniques offer the benefit that the functional relationship between the independent and forecasted variables in the evolved algorithm can be obtained by analysis, although significant effort may be required to do so.

In this paper, we present late-breaking, preliminary results wherein long-term quarterly sales of a particular, widely deployed telecommunications product were forecasted using GA's and evolutionary programming. In the nomenclature of Bäck, *et al.* (1997), the described technique would actually best be characterized as an evolutionary algorithm (EA) rather than a GA, but we will not be concerned with such distinctions. For independent variables, we elected to use readily available leading economic indicators, reasoning that these should be related to both deployment of new telecommunications infrastructure and growth of existing infrastructure. With this approach, the inherently self-organizing, unsupervised nature of GA's frees the analyst from the need to explicitly model rapidly changing nonstationary dynamics in the time series. Moreover, the method produces a deterministic model that can be studied in-depth for dependencies.

This problem is extremely challenging in general: the studies cited above were limited to short-term prediction and did not address the long-term, multistep forecasting problem treated here. The difficulty of the problem is further compounded by a lack of large quantities of sales data for use in training. Indeed, the independent variables have an inherent time resolution of months, whereas the dynamics of the sales time series can change on a time scale of less than one year. Thus, large historical data sets have limited utility in training or evolving algorithms for predicting future sales. The primary contribution of this work is to demonstrate the successful application of GA's to the long-term forecasting problem.

2. ALGORITHM DESCRIPTION

GA's were implemented using the "***e*** *Evolutionary Algorithm Program*" developed and marketed by System Dynamics International, Inc., of St. Louis, MO (System Dynamics, 1997). This software supports features of the canonical GA (Holland, 1975), as well as several extensions described below. The independent variables were 12 time series of economic indicators selected from among the many such series available on the

Economic Time Series Web Site (http://www.economagic.com). The selection of the 12 indicators used was based both on correlation studies with the actual quarterly sales data of the product of interest, and on intuitive expectations derived from extensive historical experience with this particular product.

2.1 Chromosome Structure

Each individual in the population had a single chromosome representing a LISP S-expression, or *parse tree*, comprising both operands and operations (Koza, 1992). The admissible operations were basic binary and unary calculator functions including add, subtract, multiply, divide, roots, logarithms, trigonometric transcendentals, and simple order statistics. Admissible operands included constants, up to 15 samples (perhaps causally time lagged) of one or more of the economic indicators, variables derived by combining operands via operations, and return values from *library calls*. Library calls are a sophisticated genetic feature supported by the ***e*** software used in this study. With this feature, a single gene in an individual's chromosome can represent a call to an entire chromosome from an individual in a previous population. Typically, chromosomes from the most successful individuals evolved in a previous generation are stored in a library for this purpose. Nested library calls were supported to a depth of four levels.

The maximum chromosome size for an individual was limited to 40 operations, with a corresponding number of operands based on the operator types (unary or binary). In the initial population, both the chromosome lengths and the values of individual genes were generated randomly by a uniform variable.

2.2 Fitness Evaluation

The fitness of individuals was evaluated based upon their ability to correctly predict quarterly sales data for the product. Let x_k represent the actual quarterly sales for quarter k and $\hat{x}_{k,i}$ represent the forecast of individual i for quarter k, where $k \in [1, N]$ defines the training set. We define f_i, the fitness of individual i, according to

$$f_i = \frac{1}{1+\alpha}, \tag{1}$$

where

$$\alpha = \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{x}_{k,i})^2 + \Lambda \tag{2}$$

(2)

and where $\Lambda = 10^{-5} \times$ (chromosome length) is a penalty term that favors shorter genetic programs over longer ones. We do not consider the length of the genetic program to be a particularly important factor in this application; hence the small weight given to Λ in Eq. (2) indicates that this term will be significant only in cases where two or more individuals are equally fit. When this occurs, the individual with the shortest genetic program is deemed most fit.

(1)

2.3 Reproduction and Genetic Evolution

The GA permitted individuals to reproduce both sexually and asexually. In sexual reproduction, two offspring were created with gene sequences derived from those of the parents by single-point crossover reproduction (Mitchell, 1998). A crossover point was chosen at random to divide each parent's chromosome into two gene sequences. Each offspring received one gene sequence from each parent, and these were concatenated to create the offspring's chromosome.

The population in any generation was divided into distinct ecosystems. Like mutation, the purpose of this division was to maintain population diversity and avoid trapping in local optima. The number of ecosystems was fixed at 14, the maximum supported by the ***e*** software. An equal number of individuals in the original population were randomly assigned to each ecosystem, and offspring produced in a given ecosystem remained in that ecosystem upon generation evolution. Individuals in any given ecosystem were permitted to breed with one another, but not generally with individuals in other ecosystems (except for intermarriage, as described below). Thus, competition for selection as a parent was only between individuals in a single ecosystem. This approach differs in two ways from the niches often used in multimodal optimization (Holland, 1975), (Baker, 1987), (Deb and Goldberg, 1989), (Sareni and Krähenbül, 1998). First, in the most common niching methods, individuals are grouped according to a sharing function that quantifies the similarity between individuals, whereas the ecosystems implemented here are explicitly delineated without regard to similarity. Second, with our GA all individuals in the population are evaluated using a single, common fitness function; there is no fitness sharing.

Within each ecosystem, individuals were mated as parents for sexual reproduction using roulette wheel selection (Mitchell, 1998), (Michalewicz, 1994). For a given ecosystem, let

$$F = \sum_i f_i \tag{3}$$

be the total fitness of the ecosystem. Then $\rho_i = f_i / F$ defined the fraction of the roulette wheel assigned to individual i; *viz.*, the probability that individual i would be selected as a parent in a particular sexual reproduction.

The GA also permitted sexual reproduction involving parents from different ecosystems, or *intermarriage* (System Dynamics, 1997), which occurred in each generation with a specified probability. When intermarriage occurred, roulette wheel selection was used to select parents from two randomly selected ecosystems. Single-point crossover reproduction was then used to generate two offspring, both of which were associated with the ecosystem of one of the parents.

Asexual reproduction was implemented using both mutation and inversion (Holland, 1975). In mutation, an offspring's chromosome was created by first copying the parent's chromosome and then randomly choosing a gene to be replaced with a randomly selected value of the same type (operator or operand). Likewise, inversion first copied the parent's chromosome to the offspring. Two randomly selected genes in the offspring's chromosome were then swapped. As in the case of sexual reproduction, individuals were mated as parents for asexual reproduction by roulette wheel selection. Subsequent to the production of offspring by sexual and asexual reproduction, the GA subjected offspring of both types to random mutation as described above.

After mutation, all offspring produced for a given ecosystem were placed in a new generation pool for the ecosystem. When the new generation pool for a given ecosystem

Table 1: GA parameter values

Parameter	Value
Number of Ecosystems	14
Number of individuals per ecosystem	50
Fraction of reproduction that were sexual	80%
Fraction of sexual reproductions by intermarriage	25%
Fraction of reproductions that were asexual	20%
Fraction of asexual reproductions by mutation	90%
Fraction of asexual reproductions by inversion	10%
Overall mutation rate	30%

was filled, the fitness of each individual in this pool was evaluated. A hill-climbing algorithm was then applied to the most fit individual from the pool (Michalewicz, 1994), (Mitchell, 1998). In this process, a random mutation was sequentially applied to each gene in the chromosome of the selected individual and retained only if this mutation resulted in improved fitness. Only single-pass hill-climbing mutation was applied. Multiple-pass hill-climbing produced excessive computational load with little gain in population improvement. Subsequent to hill-climbing, the new generation pool replaced the current generation of the ecosystem. Evolution continued until the stopping criterion given in Section 3 was satisfied.

3. EXPERIMENTS AND RESULTS

The actual quarterly sales time series that was used is shown in the solid line of Fig. 1, where scaling has been applied to protect the proprietary nature of the data. The forecast interval was defined to be eight quarters (two years). The time series was divided into three segments. The first segment was used for training; *i.e.*, the GA evolved populations of individuals using this segment. As this application is characterized by a paucity of data, the length of the training segment was set at N = 28 quarters. All individuals in the final generation of the GA evolution were then tested against the second, or evaluation, segment, which had a length of eight quarters. The GA was stopped when, after at least several thousand generations, only negligible improvements in fitness were observed. In all cases, this occurred after fewer than 20 thousand generations had evolved. After each run of the GA, the individual delivering the minimum mean-squared error (MSE) predictions on the evaluation segment was used to forecast the third, or test segment, which also had a length of eight quarters.

The algorithm described in Section 2 represents a reasonably sophisticated GA, for which a number of parameters such as percent sexual reproductions, percent asexual reproductions, and mutation rates must be specified. The empirically selected parameter values used in this study are given in Table 1. Note that the mutation rate was chosen quite high. We have observed the GA performance to be reasonably insensitive to the values of the parameters for this particular forecasting application.

Forecasting performance of the fittest individual produced by a typical run of the GA is also illustrated in Fig. 1. Out of the maximum of 40 operations allowed, the chromosome of this individual utilized only 10. Four of these were library calls, while the remaining six were unary and binary operations. Nested library calls were made to the maximum depth of four levels, and 18 library routines were called in total. The fore-

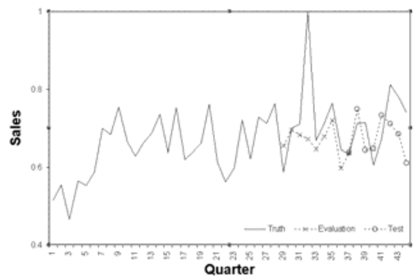

Fig. 1: Genetic algorithms forecasting results.

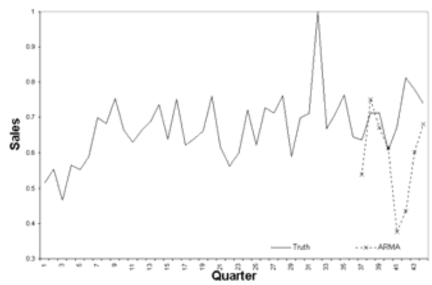

Fig. 2: ARMA forecasting results.

casting algorithm implemented by the fittest individual utilized values from 10 of the 12 available economic variables, as enumerated in Table 2.

The portion of the dashed line in Fig. 1 marked by X's depicts the fittest individual's eight-quarter forecasts over the evaluation segment. While significant error is observed at one datum, the forecasts are generally in excellent agreement with the actual sales. The fittest individual's eight-quarter forecasts on the test segment are shown by the portion of the dashed line in Fig. 1 marked by O's. For these forecasts, the root mean squared (RMS) forecast error is less than the average of sales over a two-week period, a result that we have been unable to obtain previously using more conventional techniques such as Box-Jenkins models and linear regressions.

For a typical run of the GA producing a total of 700 individuals in 14 ecosystems, the MSE and RMS forecast error averaged across the fittest individual from each ecosystem was calculated, and is given in Table 3. The average quarterly sales over the test segment are 0.7085 product units, corresponding to a two-week sales average of 0.1090 units. Thus, each of the 14 ecosystems produced at least one individual delivering eight-quarter forecasts for which the RMS forecast error was below the average two-week sales figure.

For comparison, eight-quarter forecasts were made for the test segment of the data using an ARMA(12,6) model of the type described by Masters (1995). For independent variables, the ARMA model used only the history of the sales time series and did not consider the 12 economic variables employed by the GA. Despite this difference, we feel that the results provide a fair means of comparing performance of the GA to that of a

Table 2: Economic Variables Used by Fittest Entity

Privately-Owned Housing Starts	S&P 500
Federal Nondefense Gross Investment	Bank Prime Loan Rate
Total Industrial Production Index	Commercial & Industrial Loans
Index of Leading Economic Indicators	Personal Income
Residential Fixed Investment	Federal Funds Rate

Table 3: Cross-Ecosystem Average Forecasting Error for 14 Fittest Individuals

	Max	Average	Min
MSE	0.0069	0.0065	0.0059
RMS	0.0833	0.0806	0.0770

technique that is both well established and widely used in practice. The ARMA parameters were trained over the first segment of the sales time series, and the resulting forecasts are shown in Fig. 2, where the actual sales data are also repeated. For the test segment, the ARMA MSE and RMS forecasting error were 0.0346 and 0.1860, respectively. Thus, the MSE of the ARMA model was greater than that achieved by the GA by a factor of approximately five.

4. CONCLUSION

In this paper, a genetic algorithm (GA) was used to forecast long-term quarterly sales of a product in the highly volatile communications technology industry. The independent variables input to the GA were time series containing historical values of 12 economic indicators. The GA was reasonably sophisticated, permitting sexual reproduction within and between ecosystems, asexual reproduction by both mutation and inversion, random mutation of offspring produced by both sexual and asexual reproduction, library calls to the complete chromosome of a previously evolved individual, and hill-climbing for the most fit offspring. A complete replacement strategy was used to evolve generations.

While the results obtained for this extremely difficult forecasting problem are of remarkably high quality, one should bear in mind that this research is still in a preliminary stage. Our future research will study alternative configurations of the GA, its sensitivity to parameter values such as the mutation rate and the relative rates of sexual and asexual reproduction, and improved strategies for parameter selection in forecasting applications of this type. In addition, we are hopeful that an increased understanding of the telecommunications technology business as related to specific leading economic indicators may emerge as a fortuitous byproduct of the work.

ACKNOWLEDGMENTS

This work was supported in part by a grant from Lucent Technologies, Inc.

NOMENCLATURE

F : total fitness of an ecosystem
N : number of quarters in the training set
f_i : fitness of individual i
x_k : actual quarterly sales for quarter k
$\hat{x}_{k,i}$: forecast made by individual i for quarter k
Λ : penalty term favoring shorter genetic programs over longer ones
α : penalty appearing in the denominator of the fitness f_i
ρ_i : fraction of roulette wheel assigned to individual i in a sexual reproduction

REFERENCES

Baker, J. E., 1987, "Reducing Bias and Inefficiency in the Selection Algorithm," *Proc. 2nd Int. Conf. Genetic Algorithms*, J. J. Grefensette, ed., Hillsdale, NJ, pp. 14-21.

Bäck, T., Hammel, U., and Hans-Paul, S., 1997, "Evolutionary Computation: Comments on the History and Current State," *IEEE Trans. Evol. Comput.*, Vol. 1, No. 1, pp. 3-17.

Deb, K., and Goldberg, D. E., 1989, "An Investigation of Niche and Species Formation in Genetic Function Optimization," *Proc. 3rd Int. Conf. Genetic Algorithms*, J. D. Schaffer, ed., San Mateo, CA, pp. 42-50.

Donaldson, R.G., and Kamstra, M., 1989, "Using Artificial Neural Networks to Combine Financial Forecasts," *J. Forecasting*, Vol. 5, pp. 559-583.

Goonatilake, S., Campbell, J.A., and Ahmad, N., 1994, "Genetic-Fuzzy Systems for Financial Decision Making," in *Advances in Fuzzy Logic, Neural Networks and Genetic Algorithms*, T. Furuhashi, ed., Springer-Verlag, Berlin, pp. 202-223.

Harrald, P.G. and Kamstra, M., 1997, "Evolving Artificial Neural Networks to Combine Financial Forecasts," *IEEE Trans. Evol. Comput.*, Vol. 1, No. 1, pp. 40-52.

Holland, J.H., 1975, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI.

Kim, D. and Kim, C., 1997, "Forecasting Time Series with Genetic Fuzzy Predictor Ensemble," *IEEE Trans. Fuzzy Systems*, Vol. 5, No. 4, pp. 523-535.

Koza, J.R., 1992, *Genetic Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, MA.

Kuan, C., and White, H., 1994, "Artificial Neural Networks: An Econometric Perspective," *Econometric Rev.*, Vol. 13, pp. 1-91.

Masters, T., 1995, *Neural, Novel, & Hybrid Algorithms for Time Series Prediction*, John Wiley & Sons, New York.

Meyer, T.P., and Packard, N.H., 1992, "Local Forecasting of High-Dimensional Chaotic Dynamics," *Nonlinear Modeling and Forecasting*, M. Casdagli and S. Eubank, ed., Addison-Wesley, Redwood City, CA, pp. 249-263.

Michalewicz, Z., 1994, *Genetic Algorithms + Data Structures = Evolution Programs*, 2 ed., Springer-Verlag, Berlin.

Mitchell, M., 1998, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, MA.

Packard, N.H., 1990, "A Genetic Learning Algorithm for the Analysis of Complex Data," *Complex Systems*, Vol. 4, pp. 543-572.

Saravanan, N., 1993, "Evolving Neural Networks: Application to a Prediction Problem," *Proc. 2nd Annual Conf. Evol. Prog.*, La Jolla, CA, pp. 72-78.

Sareni, B. and Krähenbül, L, 1998, "Fitness Sharing and Niching Methods Revisited," *IEEE Trans. Evol. Comput.*, Vol. 2, No. 3, pp. 97-106.

System Dynamics International, Inc., 1997, ***e** User's Manual*, St. Louis, MO.

Wan, E.A., 1994, "Time Series Prediction by Using a Connectionist Network with Internal Delay Lines," *Time Series Prediction*, A.S. Wiegand and N.A. Gershenfeld, ed., Addison-Wesley, Redwood City, CA, pp. 195-217.