Bayesian Learning with Selective Subsets of Populations in Genetic Programming

Andrew J. Novobilski, Ph.D.
Computer Science
The Univ. of Tennessee at Chattanooga
Chattanooga, TN

Farhad A. Kamangar, Ph.D.
Computer Science and Engineering
The University of Texas at Arlington
Arlington, TX 76019

ABSTRACT

An interesting element in the forecasting of time series data is the choice of a predictive model for the underlying process. Of particular interest are predictive models based on Bayesian Networks organized according to a two-tiered approach that loosely follows the human cognitive model. Past work has shown that an evolutionary search for this type of Bayesian Network in which the fittest model is chosen as the representation of the underlying process model provides reasonable results in the prediction of both synthesized and stock market data. However, observation of the evolutionary process also indicates that prior to concentrating the search in the neighborhood of one best model, several very good models are generated and then discarded over time. In this paper, the authors explore the retention of such good models and their combined use under Bayesian Learning to provide a forecast of time series data. A method for selecting individual models is presented that builds a set of forecast models over multiple generations of the genetic programming search process. Experimental results in comparing the single best model to the combined model are presented with suggestions for future study related to defining the fitness of the genetic process to evolve better predictive sets.

INTRODUCTION

An important element in the forecasting of time series data is the choice of a predictive model for the underlying process. Of particular interest to the authors are predictive models like the one shown in Figure 1 based on Bayesian Networks (Novobilski and Kamangar 2000). A Bayesian network, also known as a belief network, causal network, and influence diagram, is a graphical modeling language for representing uncertain relationships (Heckerman, Mamdani, et. al 1995). A Bayesian network is a directed acyclic graph with nodes representing the attributes of the model and directed links representing a causal relationship between parent and child. Together, this information represents the dependence between variables and gives a concise specification of the joint probability of the model (Russel and Norvig 1995). It provides a white box approach to representing relationships that exist within the domain being modeled and can handle inferencing in the absence of complete information.

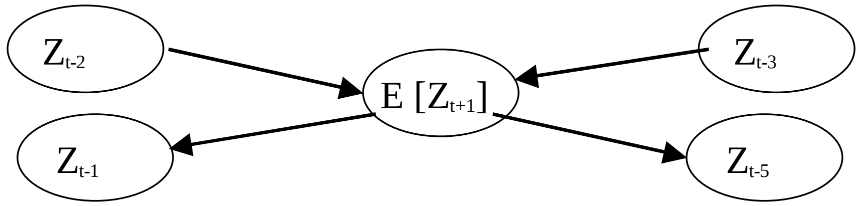

Figure 1 - A Bayesian Network indicating that the expected value of Z at time t+1 is a result of Z at time t-2 and t-3, and that the expected value at t+1 influences the value of Z at time t-1 and t-5.

Past work has shown that it is possible to employ genetic programming based search to identify Bayesian Networks to be used in forecasting future values of time series data (Novobilski and Kamangar 2000). Genetic Programming, best described as a search procedure through model space based on the mechanics of natural selection (Holland 1992) and genetic algorithms (Goldberg 1989), extends the concept of genetic algorithms by allowing a richer tree representation to be used to encode the solution being sought (Koza 1996).

The tree used by the genetic program to evolve populations of member Bayesian networks, like the one shown in Figure 1, is best described as a set of trees within a tree structure, or a tree of trees. This structure is an analog to the two-tier cognitive model (Honavar and Uhr 1989) that provides both high level decision capabilities and support for low level observations. The root tree, is used to describe the member variables contained within the network and the conditional relationships between them. Each node contains a single integer attribute (V) that will be used to determine some aspect of the network based on the node's position relative to the root. The first child node always represents d_i, while each of the remaining children represent the remainder of the Z_{t-i}'s used in defining the predicting model. A detailed description of the algorithms involved can be found in (Novobilski 2000)

The natural selection process supported by genetic programming is applied to the selection of a forecast model in the following steps:

1. Generate an initial population of candidate predictive Bayesian networks that fully specify the organization of the net work.
2. Use the same data set to train each of the candidate networks.
3. Compute the fitness of each candidate using the a second common data set. Fitness is measured as the Mean Average Percent Error (MAPE) across the test data..
4. Create a new generation by: a) selecting a certain number of individuals to be carried forward (reproduction); b) creating a certain number of new individuals by combining parts of two existing individuals to create two new individuals (crossover); and c) randomly changing aspects of the new individuals with some small probability (mutation). Choices for the participating individuals is done based on weight chance - individuals with higher fitness are chosen more often.
5. Repeat steps 2, 3, and 4 until an individual with suitable fitness is obtained.

The use of a single, best fitness measure generating Bayesian network as a predictive model reflects Heckerman's discussion on using a Bayesian Network to model an unknown process (Heckerman 1996). In addition to the ranking of individual networks in a generation, work has also been done on finding ways in which an adaptive fitness function could be used to focus the search on the most promising models (Novobilski and Kamangar 2001). Although results have been mixed (Novobilski 2000), review of the models generated at different points in the evolutionary process indicate that further study of Heckerman's second suggestion - using a set of Bayesian models to represent the unknown process is warranted.

The remaining sections in this paper address the motivation for building the set representation for the weighted predictive model from observing prior experimental data. Experimental results will focus on the forecasting of real world stock market data. Specifically, the prediction of the next day price of General Electric stock (Figure 2) given its prior daily closing values.

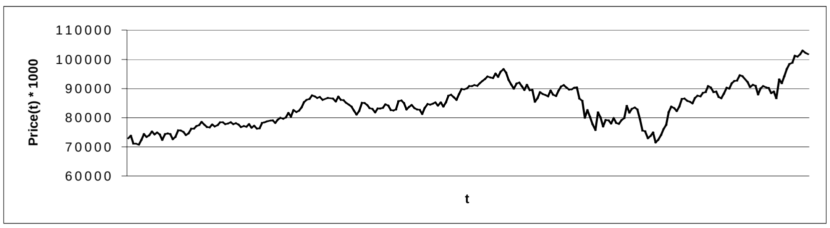

Figure 2. The Market Price of GE Stock.

The GE time series consisted of 1002 data points which were used to create a training case set of 693 records and an evaluation case set of 297 records. Each training record consisted of the target value combined with the 10 previous values of the time series. The evolutionary search process was further restricted to 10 attributes per variable selected as part of the predictive model. Finally, the population size was selected to be 50 members, with 70 generations evolved from the initial population. The evolutionary process was allowed to continue for the full 70 generations regardless of the convergence of the fitness function, worse case error or mean absolute percent error of the population members.

MOTIVATION

During analysis of the of the fitness of the best individual and overall average fitness of the entire population on a per generation basis of the stationary time series (Figure 3), the expected trend of both an improving best individual and an improving overall average fitness of the population was observed. Further analysis of the way in which the population changed (Figure 3), showed that 'new' fittest models would be introduced into the population and then begin to dominate it. A 'new' fittest model was one in which the new model differed in the lags used to forecast the expected value from the previous best model.

For example, notice the drop in the PBEST curve at generations 24, 31, 34, 42, and 67. They correspond the fact that a new individual became dominant at that time that was composed of different lags. Under the prior method of selecting a single best individual to represent the unknown model, the other two individuals would be ignored. Since the four ignored individuals had fitness rankings very close to the fittest individual, the question became one of information. Would it possible that a predictive model using a combination under Bayesian Learning of the best models that differed in the lags they used out perform a predictive model using only the single best individual?

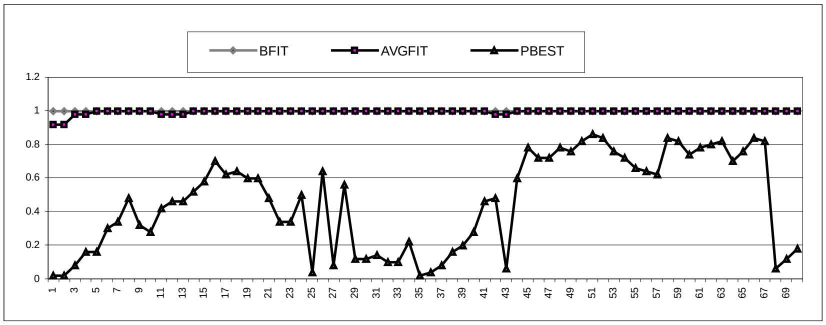

Figure 3 - Fitness of best individual (BFIT) and average fitness of all individuals (AVGFIT) during the natural selection for a predictive model for the non-stationary process. The percentage of the population that is a variation of the model that hast the best fitness.

APPROACH

In order to explore the differences in the single Bayesian network predictive model versus using a weighted set of Bayesian Network models, it was necessary to make two changes to the forecasting framework. First, it was necessary to create a means for evaluating the best fit individual in each generation to see if it should be added to the set of networks serving as the predictive model. Second it was necessary to define the probability that a member network was making an accurate forecast.

Evaluation of each generations fittest individual was done as follows:

1. Compare the candidate with the current members of the forecasting set.
2. If the candidate is "different', i.e., does not rely on the exact same lags, it is added to the forecasting set. Also, the MAPE score is recorded for the individual.
3. If the candidate is already represented in the set, then the individual with the highest fitness value is kept and the other one discarded.

Step 3 allows each member in the set to realize the advantage of natural selection in the genetic programming process by allowing a more fit refinement of an individual to replace the one already in the set.

The second adjustment to the framework involved implementing the combined Bayesian net forecast model. The forecast is computed according to:

$$E[Z_{t+1}] = \sum_{i \in M} \hat{f}_i(\vec{Z}_t) P(\hat{f}_i(\vec{Z}_t) \mid T) \quad (1)$$

where M is the set of Bayesian networks in the forecasting set, and $P(\hat{f}_i(\vec{Z}_t) | T)$ is the probability that the i^th^ Bayesian network is correct given the training data, T. This probability is calculated as a function of the MAPE for each Bayesian Network identified during the natural selection process by first summing the total of all MAPE values:

$$R_i = \frac{M_{total}}{MAPE_i} \tag{2}$$

$$M_{total} = \sum MAPE_i \tag{3}$$

and then using that value to compute the ratio R_i of M_{total} to each individual $MAPE_i$: finally, the value of $P(\hat{f}_i(\vec{Z}_t) \mid T)$ is calculated as:

$$P(\hat{f}_i(\vec{Z}_t) \mid T) = \frac{R_i}{\sum_j R_j} \tag{4}$$

RESULTS

Results were obtained for the sample data set shown in Figure 3. Table 1 provides the training results and test results for both the weighted average of multiple models and the single best model forecasts for the GE dataset shown in Figure 3. Note that WAPE measures the Worst Average Percent Error over the entire data set. It is interesting to note that although a single best Bayesian net forecast out performed the combined forecast over the training data, the combined forecast performed best over the test data.

Figures 4 shows the forecast results in relationship to the actual value. It should be noted that the forecasting framework actually specifies the expected value at time t+1 in terms of a delta to be added to the value at time t. From that perspective, the weighted forecast not only performed better from a MAPE perspective, but also from a directional perspective - i.e., is the time sequence going to move up or down.

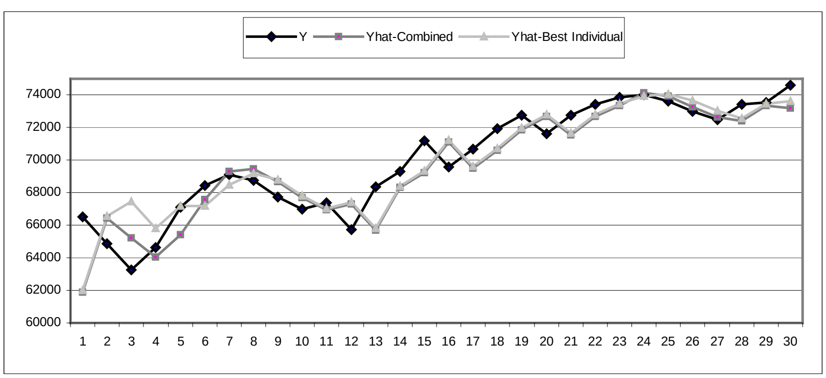

Figure 4 - The first 30 forecast values for GE dataset where Yhat1 is the weighted forecast and Yhat2 is the single best forecast.

Model	Training MAPE	WAPE	Forecast MAPE	WAPE
Weighted	0.010479	0.055562	0.013947	0.072237
Lags: 5, 8	0.010353	0.054122	0.014399	0.074125

Table 1 - Forecast results for the GE dataset.

CONCLUSION

The use of Bayesian learning to forecast time series data by combing multiple models generated by the natural selection process of genetic programming has been compared to a single best model with positive results. Although not statistically conclusive, the results do indicate the potential for the weighted forecast set to out perform a single best model when confronted with process models thought to be linear in nature.

Based on this initial work two areas for future study suggest themselves. First, in regards to the natural selection process itself, it would be interesting to see the effect of using the frequency of new model discovery to drive the natural selection process. Second, the way in which the weighting probabilities are generated during the forecast process deserves review. For this experiment, they were fixed based on the individual model's performance during training. Another possibility would be to compute the probability based on the individual model's confidence of its fore cast given the data its using to make the individual forecast. In essence, the "stronger" a model felt about its forecast, the more relative weight that value would be given during the combining phase.

REFERENCES

Novobilski, Andrew, F. Kamangar. "A Two-Tiered Cognitive Model for Forecasting Time Series Data", *2nd International Symposium on Neural Computation*, 2000.

Heckerman, D., A. Mamdani, and M. Wellman. 1995. Real World Applications of Bayesian Networks. *Communications of the ACM* 38, no. 5: 25-30.

Russell, Stuart, and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Englewood Cliffs: Prentice Hall.

Novobilski, Andrew, F. Kamangar. "Inferencing Bayesian Networks from Time Series Data Using Natural Selection", *The 13th International FLAIRS Conference*, 2000.

Holland, John. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge: MIT Press.

Goldberg, David. 1989. *Genetic Algorithms in Search of Optimization and Machine Learning*. Reading: Addison-Wesley.

Koza, John. 1996. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge: MIT Press.

Honavar, V. and Uhr, L. 1989. Brain-Structured Connectionist Networks that Perceive and Learn. *Connection Science* 1: 139-160.

Novobilski, Andrew. 2000. Forecasting of Time Series Data Using Naturally Selected Bayesian Networks. (Ann Arbor, Mich.: University Microfilms).

Heckerman, D. 1996. A Tutorial on Learning With Bayesian Networks. Redmond, WA: Microsoft Technical Report MSR-TR-95-06.

Novobilski, Andrew, F. Kamangar. "Average Percent Error Based Fitness Functions for Evolving Forecast Models." *The 14th International FLAIRS Conference*, 2001.