ESTIMATING TIME SERIES PREDICTABILITY USING GENETIC PROGRAMMING

MINGLEI DUAN
Department of Electrical and Computer Engineering
Marquette University
Milwaukee, Wisconsin

RICHARD J. POVINELLI
Department of Electrical and Computer Engineering
Marquette University
Milwaukee, Wisconsin

ABSTRACT
A new method that quantifies the genetic programming predictability of a stock's price is presented. This new method overcomes resolution and stationarity problems presented in previous approaches. A comparison, showing the advantages of the new method, is made, between the approaches, on four time series.

INTRODUCTION

Time series predictability is a measure of how well future values of a time series can be forecasted. Measuring the predictability of a time series is important because it can tell whether a time series can be predicted before making a prediction. Therefore prediction of time series with low predictability, such as a random walk time series, can be avoided. A good measure of time series predictability also provides a measure of confidence in the accuracy of a prediction. This is especially helpful to minimize the risk when making an investment decision.

After a brief background review, the previous approach in this area is introduced. Some disadvantages of this approach are then discussed, and a new modified method that aims at overcoming these disadvantages is presented and tested.

BACKGROUND

Modeling tools play an important role in estimating times series predictability. Evolutionary computation approaches provide effective tool for such modeling. These approaches include genetic algorithms (GA) [1], which are based on reproduction, recombination and selection of the fittest members in an evolving population of candidate solutions. Koza [2] extended this genetic model of learning into the space of programs and thus introduced the concept of genetic programming (GP). Each solution in the search space is represented by a genetic program. Genetic programming is now widely recognized as an effective search paradigm in many areas including artificial intelligence, databases, classification, and robotics.

There has been extensive work in the area of time series modeling using GP. Fogel and Fogel [3] added noise to data generated by the Lorenz system and the logistic map. As expected, using GP, they found that signals with no noise

are more predictable than noisy ones. Kaboudan [4] applied GP to estimate the predictability of stock price time series. The advantages of GP include its ability to evolve arbitrarily complex equations not requiring an *a priori* model, and its flexibility in selecting the terminal and function sets to fit different kinds of problems. GP has been widely recognized as an effective time series modeling method [3-6].

An η -metric was introduced by Kaboudan [6], which measures the probability that a time series is GP-predictable. By design, the computed metric should approach zero for a complex signal that is badly distorted by noise. Alternatively, the computed metric should approach one for a time series with low complexity and strongly deterministic signal.

This metric is based on comparing two outcomes: the best fit model generated from a single data set before shuffling with the best fit model from the same set after shuffling. The shuffling process is done by randomly scrambling the sequence of an observed data set using Efron's bootstrap method [7]. Specifically, the unexplained variations, which are measured by the sum of squared error (SSE) before and after shuffling of a time series $\{Y_t\}$, $t = 1,2,\cdots, N$, are compared. The unexplained variation in $\{Y_t\}$ before shuffling is

$$SSE_Y = \sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2 ,$$

where $\hat{Y}_t$ is the predicted Y_t . Shuffling increases the unexplained variation in $\{Y_t\}$ to a maximum [1]. This maximum is

$$SSE_S = \sum_{t=1}^{T} (S_t - \hat{S}_t)^2 ,$$

where $\{S_t\}$ is the shuffled $\{Y_t\}$. The measure of predictability is then defined as:

$$\eta = 1 - \frac{SSE_Y}{SSE_S} .$$

Thus, if the time series $\{Y_t\}$ is a totally deterministic signal and can be modeled perfectly, then $SSE_Y = 0$ and $\eta = 1$. If it is totally unpredictable noise, the reshuffling shouldn't affect the learned GP model accuracy, hence $SSE_Y = SSE_S$ and $\eta = 0$.

METHODS

While applying the η-metric to estimate stock price predictability, two main problems have been observed. First, the value of the metric depends on the length of the time series. Specifically, the η calculated for a 50 day stock price time series will be much larger than the η calculated from a 20 day stock price time series that is a subsequence of the 50 day series. Does this mean that a longer time series is more predictable? Of course not. In fact, there is evidence that longer stock price time series are closer to a random walk than shorter ones [5]. The source of this effect is mainly due to the nonstationarity of stock price time series. The nonstationarity becomes more evident as the sample size

increases. The second problem is a derivation of the first one. Since the η increases when the time series is longer, and its value has an upper bound of one, the value of the η-metric will be distributed in a very narrow range, especially for a long-term stock price time series. Hence, the resolution of the η-metric is reduced. This can be clearly seen by examining a long random walk time series, which has an η close to 0.9. By design it should be near zero. Since the random walk time series has very low predictability, the η-metric over all time series will be distributed in the approximate range of [0.9,1.0].

These problems are resolved as follows. For a long-term time series $\{Y_t\}$, $t = 1,2,\cdots,N$, the η-metric is calculated on the first Q points, that is, a sample series $\{Y_1, Y_2, \cdots, Y_Q\}$. Then, the sample series is shifted by τ, and the η-metric is calculated again on the new sample $\{Y_{1+\tau}, Y_{2+\tau}, \cdots, Y_{Q+\tau}\}$. Continuing this process, a series of η's that contains the local predictability estimations of subsequences of the whole time series are constructed. Generally, $\eta_{Q,t}$ can be defined as the η-metric over the sample $\{Y_{t+1}, Y_{t+1}, \cdots, Y_{t+Q}\}$. Thus, the η-series is represented by $\{\eta_{Q,0}, \eta_{Q,\tau}, \eta_{Q,2\tau}, \cdots, \eta_{Q,m\tau}, \cdots\}$. Since all the η's are estimated over same sample size Q, they are well comparable, and by selecting appropriate values of Q, they can be made to distributed in a reasonable range. This solves both problems. Additionally, by examining the resulting η-series, the variation of the predictability over time can be observed, and the overall predictability of a specific time series can be estimated by calculating the average of all η's.

EXPERIMENTS AND RESULTS

In order to test the new metric, it is applied to three different kinds of time series: a deterministic time series, a random walk time series, and two stock price time series. The experiments clearly demonstrate that different kinds of times series yield significantly different predictability results. Each SSE in the results is obtained by performing 20 GP runs and averaging the best 10.

Adil Qureshi's GPsys release 2b [8] is used to perform all the GP runs. The configuration used in this study is given in Table 1.

Table 1: GP configuration

Generations	100
Populations	2000
Function set	+, -, /, *, sin, cos, exp, sqrt, ln
Terminal set	$\{x(t-1), x(t-2), \ldots, x(t-10), R\}$
Fitness	Sum of squared error
Max depth of new individual	9
Max depth of new subtrees for mutation	7
Max depth of individuals after crossover	13
Mutation rate	0.01
Generation method	Ramped half-and-half

Deterministic Time Series

The Mackey-Glass equation is used to generate the deterministic time series in this study. The equation for the discretized map is

$$x(t+1) = x(t) + \frac{bx(t-\tau)}{1+x^c(t-\tau)} - ax(t),$$

where a=0.1, b=0.2, c=10, and τ=16. The Mackey-Glass map is seeded with 17 pseudo-random numbers and an 1100 points time series is generated. The first 1000 points are discarded to remove the initial transients. The last 100 points are used as the deterministic time series upon which the predictability metric is tested. The sample size is set to 100 for Kaboudan's method. For the new method, the sample size $Q = 20$ and the shift step $\tau = 5$. Results are shown in Table 2 and Table 3.

Table 2: Predictability of Mackey-Glass series using Kaboudan's -metric

SSE_Y	SSE_S	η
4.014×10^{-3}	4.323	0.999

Table 3: Predictability of Mackey-Glass time series using the new metric

τ	SSE_Y	SSE_S	$\eta_{20,\tau}$
0	1.938×10^{-4}	0.124	0.998
5	1.236×10^{-4}	0.089	0.999
10	6.400×10^{-4}	0.121	0.999
15	1.328×10^{-4}	0.400	1.000
20	6.691×10^{-4}	0.418	0.998
25	1.230×10^{-3}	0.254	0.995
30	6.443×10^{-4}	0.118	0.995
35	5.374×10^{-4}	0.122	0.996
40	1.009×10^{-3}	0.174	0.994
45	3.584×10^{-4}	0.100	0.996
Average η			0.997

Both Kaboudan's metric and the new metric give an average η very close to 1, indicating that the time series is highly predictable. Note that the difference in SSE_S between Kaboudan's method and the new method presented in this paper is due to the length of the respective time series. Recall for Kaboudan's method the time series is 100 observations and for the new method each subsequence is 20 observations.

Random Walk Time Series

A random walk time series is generated and tested using both the Kaboudan's η -metric and the new metric. The random walk series $\{R_t\}$, $t = 1,2,\cdots,N$, is generated by $R_t = R_{t-1} + a_t$, where a_t is random variable uniformly distributed in [-0.5, 0.5], and the initial value $R_0 = 10$. Again, for Kaboudan's method, the sample size is 100, and for the new method, the sample size $Q = 20$ and the shift step $\tau = 5$. The results are shown in table 4 and 5.

Table 4: Predictability of random walk series using Kaboudan's η-metric

SSE_Y	SSE_S	η
2.303	18.450	0.875

Table 5: Predictability of random walk series using the new metric

τ	SSE_Y	SSE_S	$\eta_{20,\tau}$
0	0.363	0.460	0.211
5	0.728	0.957	0.239
10	1.602	1.618	0.010
15	1.899	1.864	0
20	1.941	1.156	0
25	1.804	1.885	0.043
30	1.345	0.904	0
35	0.415	0.740	0.439
40	0.532	0.985	0.460
45	0.954	0.599	0
	Average η		0.140

Kaboudan's metric gives $\eta = 0.875$ for a random walk series, which is obviously not reasonable. The new metric gives an average $\eta = 0.140$, which more accurately reflects the true predictability of a random walk time series. Following Kaboudan's suggestion, if $\eta < 0$, it is simple set equal to zero, indicating that the time series is not predictable.

Stock Price Series

Next the new metric is applied to calculate the predictability of two stock price time series: Compaq Computer (CPQ) and General Electricity (GE) for the year 1999, with $Q = 20$ and $\tau = 5$. The results are shown in Table 6.

Stock Name	Average η
CPQ	0.818
GE	0.415

Table 6: Predictability estimations of stock price

The new metric gives average $\eta = 0.818$ for CPQ and $\eta = 0.485$ for GE. These η values are different from the ones we obtained from the totally deterministic time series and the random walk time series. This result suggests that the stock price series is more predictable than the random walk series, and the new metric does disclose this difference and quantifies it.

CONCLUSIONS

A new method for measuring time series predictability is proposed in this paper. It is based on the η-metric method introduced by Kaboudan [6], but overcomes the two main disadvantages of the pure η-metric method. It also provides a new feature, which shows how the predictability changes over different subsequences in a time series.

This method has been shown to be able to distinguish stock price time series and random walk time series. Future work will study a wider variety of stocks. Additionally, this method will be studied in its value in making investment decisions.

REFERENCES

[1] Holland, J. 1975. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor.

[2] Koza, John 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: The MIT Press.

[3] Fogel, D. and Fogel, L. 1996. Preliminary experiments on discriminating between chaotic signals and noise using evolutionary programming. *Genetic Programming 1996: Proceedings of the First Annual Conference.* Cambridge, MA: The MIT Press, pp. 512-520.

[4] Kaboudan, M. Genetic Programming Prediction of Stock Prices, *Computational Economics,* to appear.

[5] Chen, S-H and Yeh, C-H (1996). Genetic programming and the efficient market hypothesis. In Koza, John, Goldberg, David, Fogel, David, and Riolo, Rick (editors). *Genetic Programming 1996: Proceedings of the First Annual Conference.* Cambridge, MA: The MIT Press, pp. 45-53.

[6] Kaboudan, M. 1998. A GP approach to distinguish chaotic from noisy signals. *Genetic Programming 1998: Proceedings of the Third Annual Conference*, San Francisco. CA: Morgan Kaufmann, pp. 187-192

[7] Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

[8] Adil Qureshi's GPsys release 2b in java http://www.cs.ucl.ac.uk/staff/A.Qureshi/gpsys.html.