Volatility Forecasting using Time Series Data Mining and Evolutionary Computation Techniques

Irwin Ma
École de technologie supérieure
1100 Notre-Dame West, Montréal, (Québec), Canada, H3C 1K3
(514) 396-7687
acron@aei.ca

Tony Wong
École de technologie supérieure
1100 Notre-Dame West, Montréal, (Québec), Canada, H3C 1K3
(514) 396-8930
tony.wong@etsmtl.ca

Thiagas Sankar
École de technologie supérieure
1100 Notre-Dame West, Montréal (Québec), Canada, H3C 1K3
(514) 396-8531
thiagas.sankar@etsmtl.ca

ABSTRACT
Traditional parametric methods have limited success in estimating and forecasting the volatility of financial securities. Recent advance in evolutionary computation has provided additional tools to conduct data mining effectively. The current work applies the genetic programming in a Time Series Data Mining framework to characterize the S&P100 high frequency data in order to forecast the one step ahead integrated volatility. Results of the experiment have shown to be superior to those derived by the traditional methods.

Categories and Subject Descriptors
I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods and Search – *Heuristic Methods*; J.4 [**Computer Applications**]: Social and Behavioral Sciences – Economics

General Terms
Algorithms, Performance, Economics

Keywords
Financial Volatility, forecasting, genetic algorithm, genetic programming, data mining, S&P 100.

1. INTRODUCTION
The daily volatility is a key variable in the evaluation of financial risk and options. Integrated Volatility (IV) is calculated from the cumulative squared intraday returns of the underlying securities at high frequencies as defined by Anderson et al. [1]. In deriving IV, the daily volatility is converted from a latent variable into an observable one. In this work, the IV is evaluated based on intra-day historical data and is a more accurate approximation of the daily volatility:

$$\nu(t_m) = \nu(\Delta t, n; i) = \left[\frac{1}{n}\sum_{j=1}^{n}\left|r(\Delta t; t_{i-n+j})\right|^2\right]^{1/2},$$

where $\nu(t_m)$ is the moment rate of return distribution, Δt is the time interval of the data in which integration is done, n is the total time length of the integration and i is the total number of data.

GECCO'07, July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.

2. FORECASTING
The intraday data of S&P100 index (OEX) between 1987 and Aug. 2003 is acquired from TickData Inc. of the U. S. Part of the data set, the 15-minute high-low prices between Dec. 3, 2001 and Dec. 31, 2002 are taken for training purpose. The second part, e.g. between Jan. 2 and Aug. 29, 2003 will be used to test the validity of the rules. The first 21 days of both sets of data are used to prepare for the 21-day moving average, in order to take the monthly effect into consideration, to de-trend and to improve the forecasting accuracy. The corresponding normalized IV's were then calculated and fed into the GA's to forecast the moving directions and to find the best 100 rules by maximizing the fitness value. The GP programs are then applied to forecast the IV values at the selected time ranges ahead, e.g., one day ahead.

3. RESULTS
The fitness of the GP operation in the current investigation is derived from the Mean Absolute Error between the generated individual and the actual IV value [2].

Table 1. Forecasting accuracy for 2003 (using 2002 data)

GP generations	min	max
25	71.23%	74.4%
50	73.4%	76.77%
100	69.1%	75.4%

Table 2. Forecasting accuracy for 2003 (using 2001/2002 data)

GP generations	min	max
25	73.7%	78.3%
50	76.51%	80.20%
100	74.86%	79.1%

4. REFERENCES
[1] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P., The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, no. 96, 42-55, 2001

[2] Diggs, D. H., Povinelli, R. J., A Temporal Pattern Approach for Predicting Weekly Financial Time Series. *Artificial Neural Networks in Engineering*, 707-712, 2003