Financial Data Mining with Genetic Programming: a Survey and Look Forward

Nicolas NAVET – INRIA
France
nnavet@loria.fr

Shu-Heng CHEN – AIECON/NCCU
Taiwan
chchen@nccu.edu.tw

ISI 2007 - 08/23/2007

1

Genetic programming

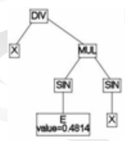
- GP is the process of evolving a population of computer programs, that are candidate solutions, according to the evolutionary principles

Generate a population of random programs

Evaluate their quality ("fitness")

Solution

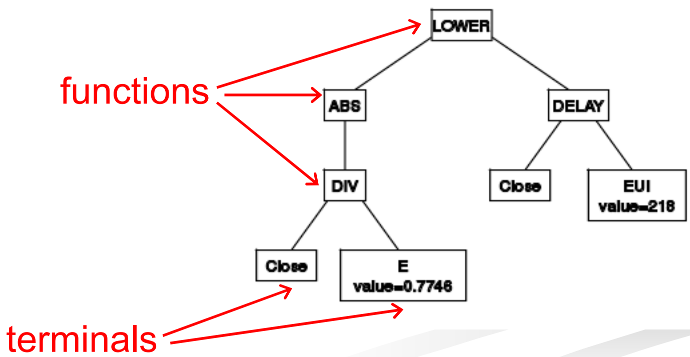
Create better programs by applying genetic operators, eg
- mutation
- combination ("crossover")



2

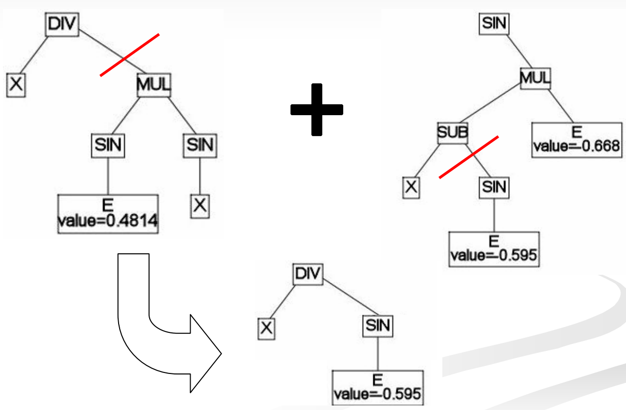
In GP, programs are represented by trees

- **Trading system:** buy if $\mathrm{abs}(\mathrm{Close}(t)/0.7748) < \mathrm{Close}(t-218)$



Typical genetic operator: standard crossover

- Standard crossover : exchange two randomly chosen sub-trees among the parents



Strong points of GP

- Solutions are produced under a symbolic form that can be analyzed by humans
- GP does not assume a predefined size and shape: it creates both the functional form and the parameters' values

➢ "Ability to produce a large number of different, yet meaningful hypotheses .. that are non-intuitive and sometimes provocative" [Kei02]

5

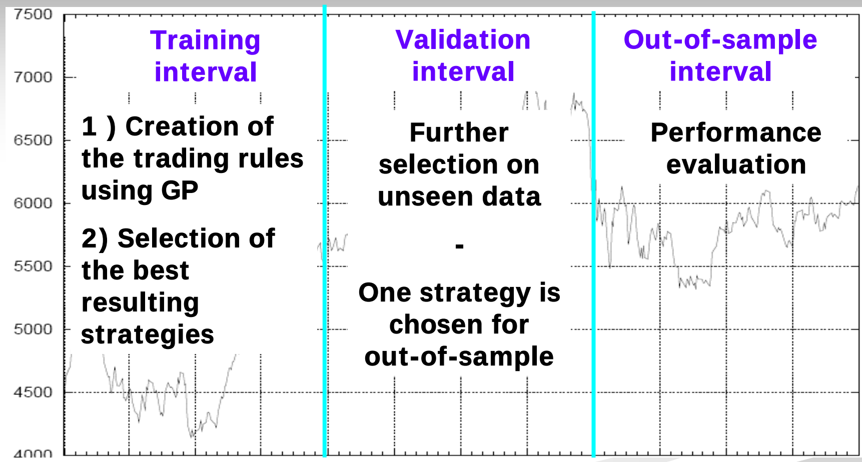
G.P. in the financial domain

1. Knowledge discovery : results are scarce
- Agent based modeling: study the evolution of a population of decision rules
- Testing the EMH in real and artificial markets

2. Financial trading :
- Composing portfolios
- Evolving structure of NN used for prediction
- Predicting price evolution
- Discovering trading rules

6

Discovering trading rules : the big picture



7

Improvements ahead of us (1/2)

1. Rigorous assessment of the GP outcomes: controlling the data-mining bias!
2. Selecting the right time series: market can be efficient
3. Reducing variability of the results from GP run to GP run
4. Re-thinking the data-division scheme for training, validation and testing periods

8

Improvements ahead of us (2/2)

5. Pre-processing the data ?!?
6. Re-thinking fitness functions : GP-friendly, sensitivity and risk adjusted, …
7. Embedding more domain specific knowledge : GP function set is still very primitive ..

1. Rigorous assessment of the GP outcomes

GP's outcomes on the training interval (1/2)

- Assume an "inefficient" solution leads to a profitable trade with probability 0.5

Probability than an inefficient system achieves a given success rate for a given number of trades

		Number of trades		
		10	50	100
Success rate	60%	0.38	0.1	0.03
	70%	0.17	$3 \cdot 10^{-3}$	$4 \cdot 10^{-5}$

➢ Guideline : penalize or discard systems with few trades

GP's outcomes on the training interval (2/2)

Probability than at least one inefficient system achieves a success rate = 70% for a given number of solutions

		Number of trades		
		10	50	100
Number of solutions tested	100	1	0.28	0.004
	1000	1	0.96	0.38
	50000	1	1	0.85

➢ NB : in a typical GP run, 50000 solutions are tested and the average number of trades is usually small …

GP's outcomes on the testing period [ChNa07]

- Compare GP with several variants of
  - Random search algorithms
    - ➔ "Zero-Intelligence Strategies" - ZIS
  - Random trading behaviors
    - ➔ "Lottery trading" - LT

Issue : how to best constrain randomness ?

- Statistical hypotheses testing
  - Null : GP does not outperform ZIS
  - Null : GP does not outperform LT

13

2. Selecting the Right Time Series

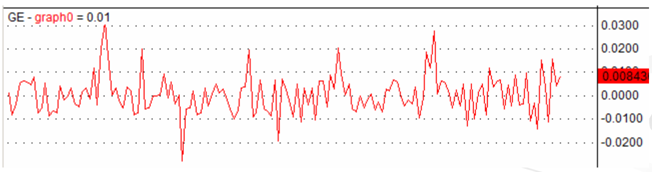
Experiments [CIEF2007]:
Does low entropy imply better profitability of GP-induced GP Trading Rules ?

NYSE US 100 Stocks
Daily Data from 2000 to 2006

14

Experimental setup

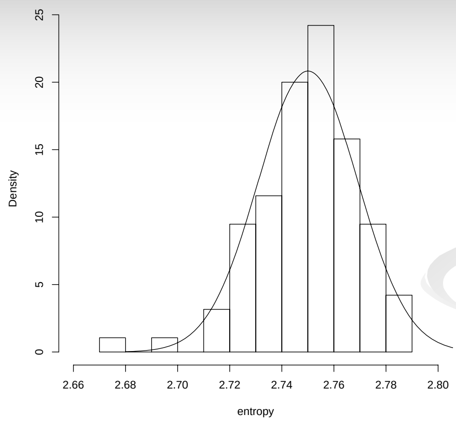
- Entropy rate estimator: Kontoyannis et al 1998
- $r_t = ln(\frac{p_t}{p_{t-1}})$
- Discretization: $\{r_t\} \in \mathbb{R} \rightarrow \{A_t\} \in \mathbb{N}$



3,4,1,0,2,6,2,...

alphabet of size 8 - equal number of values in each bin ➔ max. theoretical entropy = 3

Entropy of NYSE US 100 stocks – period 2000-2006



Mean = Median = 2.75

Max = 2.79

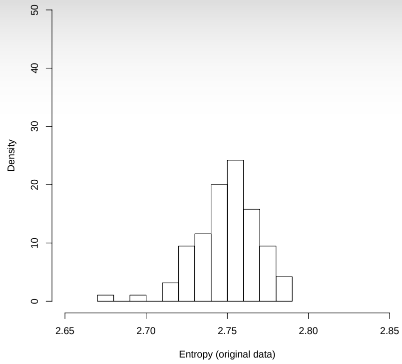
Min = 2.68

Rand() boost = 2.96

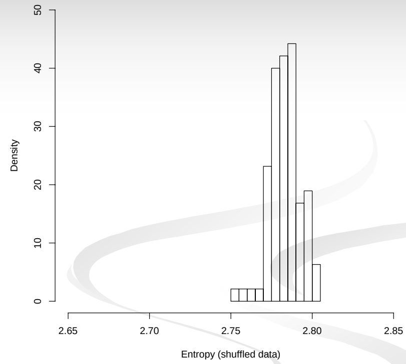
Rand() C lib = 2.77 !

NB : a normal distribution of same mean and standard deviation is plotted for comparison.

Entropy is high but price time series are not random!



Original time series



Randomly shuffled time series

Stocks in the distribution’s tails

Highest entropy time series

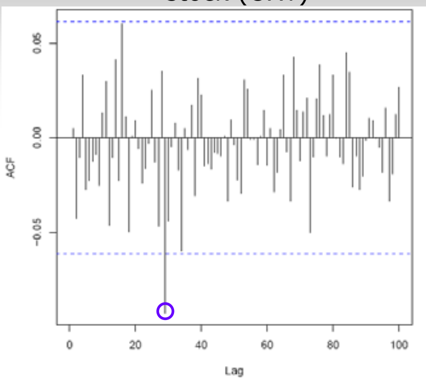
Symbol	Entropy
OXY	2.789
VLO	2.787
MRO	2.785
BAX	2.78
WAG	2.776

Lowest entropy time series

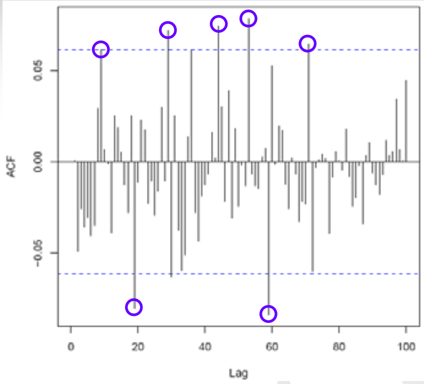
Symbol	Entropy
TWX	2.677
EMC	2.694
C	2.712
JPM	2.716
GE	2.723

Autocorrelation analysis

High complexity stock (OXY)



Low complexity stock (C)



- Up to a lag 100, there are 2.7 x more autocorrelations outside the 99% confidence bands for the lowest entropy stocks than for the highest entropy stocks

BDS tests: are daily log price changes i.i.d ?

Lowest entropy time series

m	δ	TWX	EMC	C	JPM	GE
2	1	18.06	14.21	13.9	11.82	11.67
3	1	22.67	19.54	18.76	16.46	16.34
5	1	34.18	29.17	28.12	26.80	24.21

Highest entropy time series

m	δ	OXY	VLO	MRO	BAX	WAG
2	1	5.66	4.17	6.69	8.13	7.45
3	1	6.61	5.35	9.40	11.11	8.89
5	1	9.04	6.88	13.08	15.31	11.17

- Null that log price changes are i.i.d. always rejected at 1% level but - whatever BDS parameters - rejection is much stronger for high-entropy stocks

Results: surprisingly ..

On high-entropy stocks

- GP is always profitable
- LT is never better than GP (95% confidence level)
- GP outperforms LT 2 times out of 5 (95% C.L.)

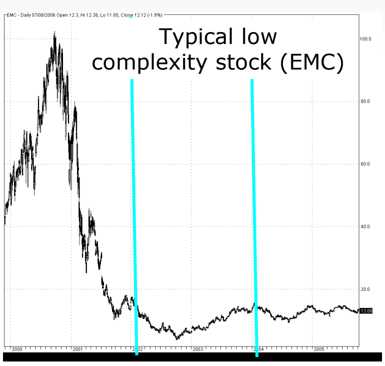
On low-entropy stocks

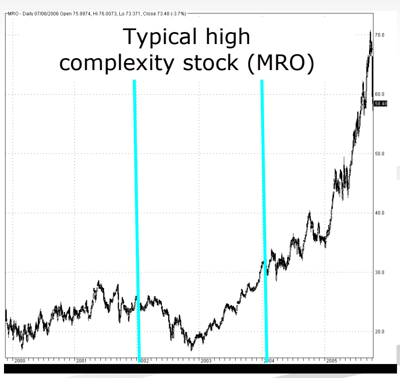
- GP is never better than LT (95% C.L.)
- LT outperforms GP 2 times out of 5 (95% C.L.)

21

Explanations (1/2)

- GP is not good when training period is very different from out-of-sample e.g.



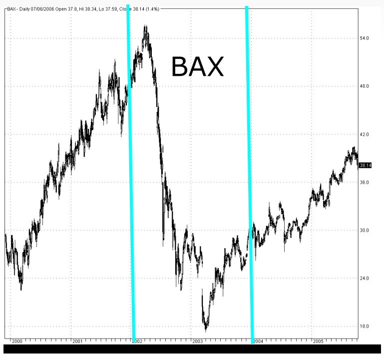


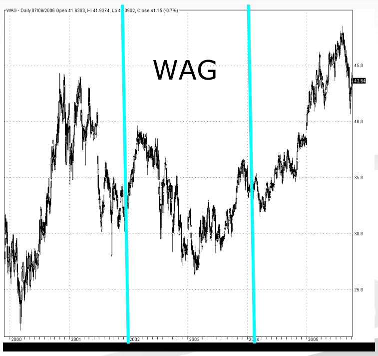
2000 2006 2000 2006

22

Explanations (2/2)

- The 2 cases where GP outperforms LT : training quite similar to out-of-sample





2000 2006 2000 2006

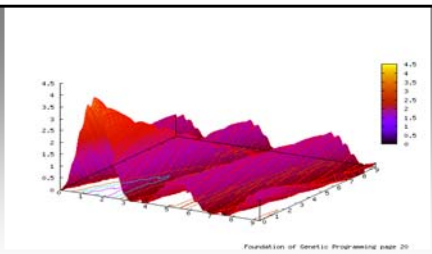
4. Re-thinking data division scheme

Data division scheme

- There is multiple evidence that GP performs poorly when training interval ≠ from the out-of-sample interval …
- What is needed: characterization of the market condition – similarity measure
- Re-learning triggered when similarity or performances below a threshold

5. Re-thinking fitness functions

Rethinking fitness functions



from [LaPo02]

- **Issue 1 :** some fitness functions induce a "difficult" landscape for GP ➔ GP-friendly fitness
- **Issue 2 :** a few lucky trades alone may lead to an outstanding return ➔ risk-adjusted fitness
- **Issue 3 :** solutions located on peaks of the fitness landscape are not robust out-of-sample ➔ sensitivity-adjusted fitness

7. Embedding more domain specific knowledge

Embedding more domain specific knowledge

- Choice of the function/terminal sets is crucial – no guidelines - 2 risks:
  - Extraneous functions
  - Required functions not available
- As yet, GP uses a very primitive language
- ➔ Enrich primitive set with volume, indexes, bid/ask spread, …
- ➔ Enrich function set with cross-correlation, predictability measure, …

29

References (1/2)

- [ChKuHo06] S.-H. Chen and T.-W. Kuo and K.-M. Hoi. "Genetic Programming and Financial Trading: How Much about "What we Know"". In 4th NTU International Conference on Economics, Finance and Accounting, April 2006.
- [ChNa06] S.-H. Chen and N. Navet. "Pretests for genetic-programming evolved trading programs : "zero-intelligence" strategies and lottery trading", Proc. ICONIP'2006, Hong-Kong, October 2006
- [ChNa07] S.-H. Chen, N. Navet, "Failure of Genetic-Programming Induced Trading Strategies: Distinguishing between Efficient Markets and Inefficient Algorithms", Chapter 8, Evolutionary Computation in Economics and Finance: Volume 2, Springer, ISBN3540728201, 2007.
- [NaCh07] N. Navet, S.-H. Chen, "Entropy rate and profitability of technical analysis: experiments on the NYSE US 100 stocks", 6th International Conference on Computational Intelligence in Economics & Finance (CIEF2007), Salt-Lake City, USA, July 2007.
- [Kab02] M. Kaboudan, "GP Forecasts of Stock Prices for Profitable Trading", Evolutionary computation in economics and finance, Kluwers, 2002.

30

References (2/2)

- [SaTe02] M. Santini, A. Tettamanzi, "Genetic Programming for Financial Series Prediction", Proceedings of EuroGP'2001, 2001.
- [BhPiZu02] S. Bhattacharyya, O. V. Pictet, G. Zumbach, "Knowledge-Intensive Genetic Discovery in Foreign Exchange Markets", IEEE Transactions on Evolutionary Computation, vol 6, n° 2, April 2002.
- [LaPo02] W.B. Langdon, R. Poli, "Fondations of Genetic Programming", Springer Verlag, 2002.
- [Kab00] M. Kaboudan, "Genetic Programming Prediction of Stock Prices", Computational Economics, vol16, 2000.
- [Wag03] L. Wagman, "Stock Portfolio Evaluation: An Application of Genetic-Programming-Based Technical Analysis", Genetic Algorithms and Genetic Programming at Stanford 2003, 2003.
- [Dem05] I. Dempsey, "Constant Generation for the Financial Domain using Grammatical Evolution", Proceedings of the 2005 workshops on Genetic and evolutionary computation 2005, pp 350 – 353, Washington, June 25 - 26, 2005.
- [Kei02] M. Keijzer, "Scientific discovery using Genetic Programming", Phd Thesis, DTU, Lyngby, Denmark, 2002.