Peter Shepherd. 26th May 2008.

MSc Intelligent Web Technology.

B39RB3 Research Methods portfolio.

"Web Pattern Recognition in News Providers and Financial Data."

Supervisor: Professor David Corne.

I, Peter Shepherd, confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the words of other authors in any form e.g., ideas, equations, figures, text, tables, programs etc. are properly acknowledged. A list of references employed is included.

_____26th May 2008.

ABSTRACT:

This portfolio proposes and frames a 3 month dissertation project in the application of Genetic Programming to financial prediction using traditional numeric data inputs, but with additional inputs from relevant web text-mined material which has been qualitatively assessed and quantified. This latter material may also be used to guide GP rules toward selective use in expert-predicted market conditions which match those in which they were learned. This adds the dimension of machine measurement of expert reaction to real-world events being used to pre-empt diffusion of reaction into the market, to the existing element of discovering predictive data patterns too complex for routine human analysis.

The project's aims and objectives are set out before a more detailed rationale and description, and an early caste of what the final dissertation structure may look like. A selective literature review forms the bulk of the portfolio and seeks to appraise the current state of research and applications in this field and its shortfalls. It includes work already attempted around the project's specific focus and forms a base of information from which to build the basic GP software and data set-up, encompassing the major issues and difficulties in the technique: overfitting, overcomplexity, ongoing dynamic learning, and design of performance and fitness measures.

The conclusion points to the next step in the work and points out the risks and issues involved, with contingency plan for failures in the initial plan.

Table of Contents

1) Project aims and objectives.

The aim of this project is to develop a working software system in Java using Genetic Programming for financial market prediction using traditional numeric data inputs from the web. This will be used as base for a second application which will add to the first additional inputs from relevant, qualitatively assessed and quantified web text.

The principal aim is to measure any added predictive value of the second model and provide statistical evidence of its effectiveness or otherwise.

Text derived data can be used in 1 of 3 ways:

- to guide the GP to apply rules to broad, or local, market or sector conditions which are similar to the ones in which the rule was learnt as judged by online activity and sentiment
- to detect real-world events which can trigger trading rules in the minutes (or possibly hours) before the market reacts (as with current work on text data only in the literature review).
- in the simplest (fall back) case to provide informative, probability-based alerts to users from text feeds, presented parallel with the probabilities of numerically based trading signals.

Main objectives are:

1. Completion of basic GP application.
2. Completion of second GP application with text-derived data input.
3. Achievement of a range of test results measuring performance and gains of each system, against selected online training and testing data, compared to a chosen base measure of stock market performance.
4. Comparative statistical analysis of both sets of results to assess the level of statistical significance of the learning of each.

2) Project rationale and description.

Rationale of target achievement:

The project aims to achieve the contribution of evidence for or against the hypothesis that textual information sources on the web can provide significant guidance and additional learning data to Genetic Programming(GP) applications in computational finance.

Some recent existing work, detailed in the literature review, lends weight to the proposition that inputs from relevant web text-mined material which has been qualitatively assessed and quantified, may supplement and improve financial prediction using traditional numeric financial data inputs.

The literature review also shows that a core difficulty in the field of GP predictive learning in finance is that it is particularly difficult to achieve a situation where the specific conditions in the market or in the sector during the application of the rules is in any way similar to those from which the learning data was drawn. Related to this is the fact that the predictive effect of rules tends to deteriorate the longer in time after training they are applied.

This field of GP has had little in the way of conclusive success in its decade or so of research efforts. As discussed in the literature review, there are several reasons for this and for its lack of cohesive direction or consensus. As illustrated in the "Background" section it is probable that a step forward in the field is most likely to come (as so often in A.I.) from fresh perspectives or supporting techniques from closely related fields, from hybridising of machine intelligence and human judgement and expertise, or from external academic fields.

Using real-time web postings of human expertise (news, specialist message boards, expert analysis and commentary sites and blogs) can automatically elicit immediate warnings and sentiment on market-influencing events. This can add the dimension of machine measurement of expert reaction to real-world events being used to pre-empt diffusion of reaction into the market, to the existing element of discovering predictive data patterns too complex for routine human analysis.

Whether the posters of the text are skilled predictors of events, or fast reactors to events, or their activities themselves create an effect, or all three, it seems possible that GP algorithms could learn the relationship of patterns in this activity and patterns in trading activity.

Simply categorised counts of the time series of web activity can be fed into existing GP learning with prices, trading volumes and pre-calculated financial statistics. But it can also be used to guide GP learned rules toward selective application in expert-predicted market or sector conditions which match those in which they are known to have been learned, rather than attempt to learn rules that are universally applicable from a short test period and apply them forever. At the simplest it might be used to apply ordinary GP trading rules only to stocks which are in a state of short term change as judged by web comment hits, as GP research repeatedly shows that rules tend to test as successful predominantly in volatile or negative market conditions.

This project will therefore attempt to set up two sets of experimental GP runs over a selected set of online historic financial data which has an intermatching time series of both numeric figures and of web text comment. It is hoped that this may be supplied pre-classified by sentiment by proprietary online software. It will attempt to compare one set of results gained by using an ordinary GP application with no text data contribution, and one with, and compare both to some base measure of performance in order provide evidence of whether such an approach is worth further investigation.

Work toward objectives:

Prior to beginning work on the software, core decisions on practicalities have to be made in a detailed planning stage regarding:

- sources of quality text data streams, their time series length and continuity, their costs, their level of prior categorisation, and their correspondence to numeric time series
- the practicalities of the level of GP sophistication and borrowing of recent innovation (see literature review) which can be attempted in the development time calculated as available for each element

The risks and issues in the project are discussed in the conclusion.

The project objectives are all software development based (including the statistical analysis of results using spreadsheets.)

The following textbooks will be used to construct, in Java, the basic web mining software and the GP algorithms. The complex of developmental parameter choices to be made for the latter will also be guided by the literature review. The last 3 books are specifically Java oriented.

CHEN S.H., KUO T.W. AND SHIEH Y.P. (2002) Genetic Programming: A Tutorial with the Software Simple GP. in CHEN, S. H. ed. (2002) *Genetic Algorithms and Genetic Programming in Computational Finance*, Kluwer Academic Publishers,Dordrecht.

FREITAS, A. A. (2002) *Data mining and knowledge discovery with evolutionary algorithms,* Berlin ; London, Springer.

LANGDON, W. B., POLI, R. (2002) *Foundations of genetic programming,* Berlin ; London, Springer.

LANGDON, W. B., POLI, R., MCPHEE, N.B., KOZA, J.R. (contrib.) (2008) *A Field Guide to Genetic Programming,* http://www.lulu.com/items/volume_63/2167000/2167025/2/print/book.pdf

LOTON, T. (2002) *Web Content Mining with Java,* Wiley.

WITTEN, A.H, FRANK, E., (2005) *Data Mining:Practical Machine Learning Tools and Techniques with Java Implementations,* Morgan Kaufmann.

An outline description of the project and the work to be done is given in the next section.

3) Dissertation structure outline:

Abstract
Acknowledgement
Plagiarism statement
Contents table
Figures table

1 Introduction.

1.1 Aims and objectives.

- Development, according to events, of aims and objectives presented in this document.

1.2 Background.

- Short preview summary of literature review section, related more directly to the ultimate direction of the project.

1.3 Current shortfalls of GP in Computational Finance.

- Short summary of this, from literature review section.

1.4 Hypothesis and Summary of argument.

- Specific hypothesis and short dissertation argument developed from the "Rationale and Description" presented in this portfolio.

2 Literature Review.

- Similar headings and content to literature review presented in this document with revisions based on experience and any further reading. Related more directly to the ultimate direction of the project.
- Additional section relating the ultimate software development decisions and choices that will have been made, directly to the sources of the ideas, and relating the ultimate directions of the project work, to work directly preceding or suggesting it.

3 Experimental approach.

3.1 Software Design.

- In separate subsections, for each of the 2 applications, detailed description of:
    1. Design and parameter options available -- from text books and from literature review.
    2. Rationale of the individual choices made from among these options.A subsection for each aspect of GP considered, and the design features included (or rejected) and why.
    3. Summary of the final design structure and features of applications.

3.2 Software Implementation.

- Discussion of practical aspects of getting software up and running, any difficulties, technical details, choices made and reasoning, and impact of all these things on project.

3.3 Experimental Structure and Data.

- In separate subsections, for each of the 2 applications, detailed description of:
    1. Choices of data sets (types) available -- from online searching and from literature review.
    2. Rationale of the individual choices made from among these options.
    3. Summary of the structure and features of the chosen data sets.

- Detailing of the performance and fitness measures to be calculated and used for comparisons in the experiment and reasoning for the base-measures chosen.

- Detailed discussion of the structure and nature of the experiments to be run:
    1. Seeking to compare what to what and reasoning for this.
    2. Relating to the hypothesis/objectives what the rationale is for the experiments and the results sought.
    3. Details and rationale of structure of software testing runs....numbers of separate runs, use of average/best result etc.

4 Evaluation of Results.

4.1 Presentation and description of results.

- Tabular and diagramatic/graphical presentation of 3 experiment sets of results and in comparison to each other (base performance measure, with text data and without.)

- Description of relationship between them, and salient features.

4.2 Statistical Analysis and interpretation of the results.

- Discussion of the level of significance of the results (or otherwise) using statistical analysis, null-hypothesis/p-value analysis, or others suggested by reading or advised by supervisor.

- Discussion of meaning and importance of level of evidence for, or against, dissertation hypothesis.

5 Conclusions.

5.1 Achievement of Aims and Objectives.

- Or not. Against list of original aims and objectives.
- Subsection on failures, lessons learnt, what could have been done better.

5.2 Summary of Contribution.

- Compared to preceding and contributing work, and judged against project rationales in this document.

5.3 Future work.

- As suggested and conceived during work on the project.

6 Appendices.

6.1 Bibliography and references.

6.2 Program listings - main GP element.

6.3 Program listings – GP with text input.

6.4 Tables of experimental results.

- In more detail.

4) Literature Review and Background.

Bibliography and references are at the end of this section.

Background:

In a book review of "*Genetic Algorithms and Genetic Programming in Computational Finance"(2002),* [Chattoe,2004] provides an illuminating external view and critique of the eponymous field of endeavour. Some of the criticisms provide pointers towards the flaws or gaps in current work that may provide research of interest, and simultaneously the appreciable problems for new research. From the viewpoint of a sociologist, much is pointed out which may not be greatly discussed in the internal literature.

Principally there is the considerable fragmentation within the field :

> *"each (apparently arbitrarily designed) programme is being tested on a different group of data sets and using a different standard for comparison (vanilla GP, "Buy and Hold" – surely an incredibly easy standard to beat, GARCH model). Furthermore, each paper seems concerned with different aspects of GP design: one with epistasis, another with proper out-of-sample testing, a third with the effect of problem representation and a fourth with the choice of fitness function. The net effect seems to be that all the results presented are "rather better" than whatever they are being compared with but that none are overwhelmingly impressive"* [Chattoe,2004]

To the list of concerns of individual papers are added data pre-processing, validation against overfitting, and simplicity and human interprebility of rules, with each new experiment concentrating on their own concern while simplifying the others as much as possible. These are:

> *"all selected on improved fit for a (typically) small number of data sets rather than in competition with other programmes or for general learning ability..... that each paper chooses its own baseline comparison [causes]the absence of "bottom line" effectiveness results that can be used across contributions to guide architecture choice."*[Chattoe,2004]

He suggests, therefore, what will be the fall-back contingency focus of this project (see portfolio conclusion) : "it would be interesting to see how the evolutionary algorithms perform against a wider class of machine learning approaches."

One possible explanation for fragmentation hinted at is motivation toward (patentable) originality, rather than building on what is already present, and the possibility that the most successful originality may drop out of the academic pool as commercial patents.

The inherent difficulties of the field are discussed. One is its reflexivity, in that any "predictive regularities" are likely already to have been spotted and processed by the "efficient market" of economists and therefore to have been cancelled out. The reflexive circle is likely to turn on any considerable GP predictive success: like alchemy, success depletes the value of the goal.

He criticises failure to model, analyse and explore what it is that is being predicted from a sociological perspective, which is in essence "the behaviour of a population of 'somewhat

socialised' heterogeneous actors" (the community of trained economists and analysts),and the socialisation and trend-following involved in their investment decisions.

All of these concerns highlight the possibilities of the first focus of this project. This is that GP might use as additional predictive input, data comprising qualitative and quantitative appraisal of the feelings and behaviour of this community, in its reaction to real-world events; news flows, message boards, blogs, or commentator analyses. From this it might create predictions either before human actors have framed their investment reactions or before the full sociological diffusion of reactive behaviour has completed.

This pre-emptive, event-based prediction (exploiting machine reaction speed) is somewhat different in principle to the complex pattern machine learning in an immense search space which is currently the focus of GP financial applications. It is possible that adding this second string might overcome some difficulties previously listed. Particular amongst these is the suspicion that immense teams of commercial "technical analysts"(TA) are likely to have spotted, exploited and thus removed the predictiveness of most of the patterns ever present. Using the reactions, analysis and behaviour themselves of economists and investors may be a useful supplement.

The possibilities come full, reflexive circle again with Chattoe who is positive about the book contribution of [Thomas, 2002] (covered in Literature Review) on message board volumes as predictors to add to traditional variables. "However, the authors don't consider what might happen to this predictor once the discussants realise that their conversations are being used in this way!"[Chattoe,2004]

Terms:

The use of terms in this review are as follows. Unless otherwise stated "benchmark" refers to the "buy and hold" stock trading strategy of simply buying and keeping all the stocks in the test for the period of the test and "excess gains" means over this measure. Evolutionary algorithms(EA) is a set containing genetic programming(GP) and genetic algorithms(GA), among others. Genetic Network Programming is a hybrid of GP and GA which presents results in the form of graphical networks (GNP) and grammatical evolution (GE) creates GP output by mapping the output of a GA process to a Backus-Naur form grammar. Predictive financial applications using GA and neural networks have not been covered here. Technical analysis (TA) which analyses numerical patterns in the past movement of stock prices is very different from the massively predominant "fundamental analysis" which analyses market as you would expect, through financial, economic and business information. The Efficient Market Hypothesis (EMH) is traditional economics theory which stipulates that TA cannot succeed as there are no predictive patterns in the data as all information is efficiently processed by the market, setting the "correct" price. This is an intense and unresolved controversy in the economics literature. Risk-adjustment of the performance measure is done by some who feel the true value of a return can only be measured in relation to the risk taken in gaining it. Some of the parameter choices made in a GP application are abbreviated here as mutation rate (mut), crossover rate (cross), initial population size (pop), generations in the run (gen), cut off generations if no change in best rules for *n* generations (cut).

Since they are all plural authors, references are referred to in the plural.

One of the leading researchers in the field, A.Brabazon, is co-author of [Dempsey,2002], [Dempsey,2006], and [Yin,2007].

Literature Review:

> *"GP is especially useful in data rich environments; where the search space is large and highly complex; where conventional mathematical analysis cannot provide analytic solutions; and where the interrelationships among the relevant variables are poorly understood."*-[Yin,2007]

Contrary to Chattoe, Yin et al make clear that GP can achieve what human TA cannot achieve because of the sheer size and complexity of the search space - in the length of the data time series, and the richness, depth and non-linear complexity of the data, and the choice set of variables available. There are a huge number of parameter choices that the human GP modeller can make for the programme itself, usually by trial and error, and usually by choosing complexity in one area and simplicity in the rest since each one can increase processing geometrically. If these are added to the equation then the fragmentation, incohesive lack of direction, inconsistency, and marginal success in GP financial applications research mentioned by Chattoe are understandable.

It is to be noted that, though most of the papers here reference [Allen, 1999] , and authors in EA more generally before 1999, such as Koza, there is little cross reference network between the authors discussed here to be referred to.

A review of the literature of the field is essential to gain a mental list of the different choices and possibilities in the area of what others have used in terms of:

- specific markets, sectors or stocks
- specific time periods
- comparison tests and performance measures
- GP operators and parameters

Some of these choices must be made interactively during the software and data design phases of the project. Several will need to be made in the first, planning stage, including the list of more strategic choices presented by the literature.

Strategic choices concern methods to conserve the currency of the rules (against degeneration of the learning over time) and, most controversially, strategies against the considerable GP problem of overfitting. This is done either through validation methods or through simplicity and human interprebility of rules. As can be seen advocates of this last approach are somewhat in contradiction of Yin's opening quote and those who seek to preserve the full (over-efficient?) power of GP. It may well be asked why use GP if simplicity rather than complexity is the goal.

However such choices can only be made (in these papers and in this project) in the context of the time and resources available.

Base Examples:

[Allen,1999] provide the most thorough descriptive introduction to the mechanics of GP in analysing S&P500 data for 1928-1995.

Validation against a 2 year selection data period to avoid overfitting is used. It takes the *single* rule with highest excess returns in a contiguous 5 year training data period, selecting it as the new chosen rule only if it improves on the existing best rule over the selection data.

Pop=500, gen=50, cut=25, with arbitrary tree maximums of 100 nodes ,10 levels.

This is the single paper which is very frequently referred to by later peers. Despite being one of the most respected and thorough of experiments in the field, used as a point of departure by many, most are attempting to explain its negative results and conclusions.

There are many explanations, which illustrate the complexities of framing an experiment in this area. [Allen,1999] themselves conclude this is a "base case" with a relatively simple GP "which could be developed further", its data inputs are limited and simple, parameters are not necessarily optimised, particularly not for the very varied market conditions of varying complexity contained in the huge testing periods, (the longest being 1936-1995!). They continue: "it would be interesting to apply a similar technique to learn fundamental trading rules by changing the building blocks to include the desired fundamental variables." Large numbers of trades eat away profits with charges. There is no shorting of stocks due the expense of shorting a complex broad index, whereas later studies have shown most GP profits come from shorting in a volatile market.

[Potvin,2004],[Pavlidis,2007], and [Thomas,1999] also all provide a brief but thorough guide to constructing GP algorithms (with reasoning for the various parametric and structural choices) and a short history of the controversy in economics of the past 40 years between EMH and TA, listing several notable successes in the latter.

[Potvin,2004] are optimistic about GP generated rules for TA market timing. They provide the useful example of evolving rules for individual stocks (1 each of the major Canadian companies in each of 14 business sectors 1992-2000) rather than the usual broader index, allowing much cheaper and easier shorting of stocks, more sensitive individual rules, and insight into which sectors might be more receptive.

Simple raw data is used (price points and transaction volumes).

Introducing the rejecting of a new training best rule if it reduces performance over a control period by 25%+ or for 3+ generations is an interesting experiment against overfitting but appears to fail. It reduces the number of generations to create rules but doesn't improve performance at all. It is of interest that experimenting with training days/testing days split of 256/256 and 1498/256 shows the former displaying dramatic overfitting but only slightly inferior results. Decision trees after simplification are around 8 deep and 30 nodes.

From results achieved the paper concludes that evolved trading rules work well if the market is stable or falling (the benchmark buy-and-hold achieving -40% to c.7%) , not when it is rising (buy-and-hold +7%to+80%): a common finding, but the scope of the trials here does not seem broad enough. Relevant to this project is that this "indicates an appropriate context or 'timing' for triggering the application of technical trading rules".

The appreciable difference in results between companies (ie.sectors) is not usefully explored despite revealing the fact that if the worst performing 2 stocks were taken out by somehow identifying that the macro market conditions for the sector were very different in the training and testing periods, then overall excess returns would be almost 10%, rather than almost -5%.

[Pavlidis,2007] provide a different angle of comparison for GP trading rules, comparing performance against Generalized Moving Averages, a traditional trading rule used in TA. The latter are here optimised with Differential Evolution, a simplified form of evolutionary algorithm which varies only the the values of variables in a set decision structure. The performance measures in this simulation case ***are*** risk–adjusted but not by the Sharpe-ratio, which is criticised.

$/yen foreign exchange: 5292 daily observations over 1985-2007 provide the experimental data. Though the moving averages rules "proved to be more robust" what was found to distinguish GP was an ability to find much more profitable single rules and detect "novel patterns", rules with

greater accuracy longevity, and greater profitability overall. Analysis of statistical significance measures also showed them to be less suspect of being " attributable to well–known properties of the data" than moving averages.

Simplicity and Overfitting:

To address the issues of comprehensible simplicity and reduced overfitting [Becker,2003] refers back to much earlier work on decision tree pruning, suggesting that accuracy can in fact be improved in the process. Occam is cited in favour of the improved prediction of simple models that reduce generalisation error and thus trials are proposed in small numbers for GP generations and populations. That this slashes the number of models considered is seen as positive. The mechanisms designed for GP complexity reduction are a weighting against complexity in the fitness function and the use of ready-calculated technical indicators used by experts in technical trading. Though this pre-packaging admittedly biases the search it also adds expert domain knowledge. Experiments are upon S&P500 1990-2002 and "the performance of these learned rules can exceed that of a buy-and-hold strategy" in "several" cases. This seems light evidence to conclude "we can produce comprehensible rules and avoid overfitting".

In a second paper [Becker,2003/2] add to the earlier experiments, further increasing simplicity and overfitting precautions. A useful discussion is provided of the disappointing results of [Allen, 1999]. Here the latter's number of trades are reduced by 80% and in-the-market rather than out is increased to 93% instead of 57%. This simplification is achieved using monthly not daily data, and by increasing the number of derived technical indicators (for example 2,3,6,10 month moving averages, 3,10 month rate of change and various trend line indicators and price resistance markers of moving average minima and maxima) and by slashing the number of operators to just AND,OR,NOT,<,>.

Their pop=500 and gen=100 and replacement at 50% are typical. Like [Allen,1999], *S&P500* data is used, training on data from 1960-1990 and testing on data from 1991-2002. (See [Allen, 1999] review for their approach).

In the first experiment (from the first paper) average tree node and depth counts are reduced from 343 and 42 to 10 and 4 ([Allen,1999] simply had a cut-off of 100 and 10). More convincing evidence of overfitting avoidance comes here with old methods outperforming new by over a third in-sample but this result being almost fully reversed out-of-sample. But buy-and-hold is only marginally beaten by the improved version, and not at all by the old methods.

The second experiment instead uses a novel fitness function which considers the number of positive 12 month period performances ("consistency"), as well as total return, achieves much better excess return (almost 18%).

Finally the innovation of evolving rules for buying and those for selling as separated species is tested, achieving excess of over 19%.

Overfitting of test data and the opaque over-complexity of evolved programs that may contribute to it are the concerns of [Thomas,1999]. They conduct predictive GP experiments in the $/Deutsch Mark and $/yen foreign exchange markets. Results are fairly poor from the point of view of market prediction (a signal failure with the DM). Again it might be noted that the more negative and unstable market (yen) provided the success.

However the principal aim was to explore the effects on overfitting of validation and arbitrarily imposed simplicity. The experiments with validation have no positive results at all. This, along with the market prediction failure, seems likely to be because of under-fitting of the data, particularly that caused by the program "simplicity" they seek. This is not clearly pointed out.

Simplicity is imposed with an imposed cut-off depth for the decision tree of 5, 3 or 2. Though a depth of 2 gives around twice the performance of 5, the marked improvement of 5 over 3 is not adequately dealt with in the conclusion that simplifying GP rules to 2 level trees is beneficial. It certainly steers clear of overfitting.

In fairness their conclusion debates the point whether this draws into question the point of GP in this field if it is best used without any of its potential analytic power. The answer to this is conjectured as perhaps replacing the crude cut-off mechanism with a fitness mechanism designed more subtly to favour simpler trees.

[Chen,2003] offer an opposing view to the problem of overfitting of test data, discussing its avoidance through validation, which dates from the cited recommendations of [Allen,1999] and [Thomas,1999]. The procedure is summarised : a training set of data is used to construct GP programs from which either best model(s), or all models, are further selected by performance over a second selection/validation set of data, the iterative process ending, cutting off the search, where no new "best" has been found within, say, 25 generations.

The paper attempts to systematically quantify overfitting, calculating as a measure a "signal ratio" for any learned rule. This is derived as a complex statistical function of its variance, characterising the levels of noise and of signal present in the rule. Using this measure experimental assessment of current practices is conducted. Varying search intensities (that is, choice of populations and generations), both with and without overfitting avoidance techniques, they conclude that usual practice in both regards is in far greater peril of under-fitting than much feared overfitting. It is pointed out that some applications have found out-of-sample performance superior to in-sample. They suggest that the expressive power of GP to exploit the massive search space and hidden predictive information provided by financial data is therefore being much underexploited. Their simulation results suggest that the existing validation design dominant in the field is not effective against real overfitting dangers whilst in fact causing underfitting, possibly explaining the under-performance in the earlier work cited. Experimentation is recommended with combination effects of higher population and generation counts in tandem (rather than varying these individually) and in new designs for guards against overfitting.

Dynamic Learning:

In a S&P500 study through 1991-97 [Dempsey,2002] investigate the core issue of the degradation over time of efficacy of the rules evolved by dividing their out-of-sample 2190 testing days (after 440 training days) into 6 equal tranches over time. This also allows their assessment in differing market conditions.

Importantly for this project they touch on past TA work which evidences impurity of the market model and learnable patterns in relation to days of the week, seasons, news and analyst coverage, sheer momentum, and market reactions which are late, excessive or the reverse. Recent economic studies suggesting a temporary phase of analysable market impurities is now receding are countered by the continued size and cost of TA departments.

Moving Average indicators are chosen for analysis here from among the widely used myriad available, such as indicators of Momentum, Trading range, and Oscilators. They reject the inclusion of risk weighting of returns in earlier studies' fitness measure as "risk should be already factored into the trading strategy", but nevertheless the experimental comparison to buy-and-hold is negative and defensively talked around, listing successes in the past against the FTSE100, ISEQ, DAX, and Nikkei and citing overfitting to diverse market conditions over the training data, but also the strong upward trend in the test period.

However the central experimental comparison, which shows some marginal success, is

between their standard GE methodology and GE with a novel constant evolution process which is a complex function applied to the GA portion of their GE process (therefore beyond the scope of this review). Another aspect of interest allows a buy signal's monetary size to be determined by the size of the signal divided by the previous largest buy signal (sell signals remain absolute). In such poor test conditions though, the impression is that proper evaluation has not occurred.

[Dempsey,2006] again use GE and cash positions varied by signal strength whilst revisiting the deterioration of rule efficacy with longevity, and through dynamically changing market conditions. An adaptive trading system is developed using a "moving window" to continue training new rules during system life but including attempts to maintain a memory of past good rules. The latter is achieved by simply using the end population of the previous retraining as the start population of the next retraining on the fresh window of data resulting from moving the window on by the next time increment. Though not attempted here (fixed at 10), adaptive variation of the generations parameter at each retraining can vary the amount of "memory" versus adaptation, as a function of the level of change measured between the data time windows.

In the comparison to the performance gained from an ordinary GE algorithm - which restarts with a fresh population at each time increment - and the methodology of "using a single adaptive population across the time series" the latter performs considerably better and avoids the problem of the former which generates overly sensitive (overfitted) rules.

Future work recommendations are experiments in varying the size of each time increment (a more adaptive system, or less) and the window width (watching out for data snooping) and relating performance/"memory" achieved in each time increment's population to its evolution generation count.

In a reverse of policy,Sharpe ratios are used here (returns/their volatility) to risk adjust performance. It is remarkable that they are pleased with the system performance (122%) against the benchmark index (133%) - S&P500 1992-97 – as, they say, this was a rising market allowing little shorting! But applying the new system only to Nikkei225 1993-97 – a volatile market – gets results 74% better than the index, largely through shorting it appears.

[Chen,2007] have 2 papers describing GNP combined with additional contributions from Reinforcement Learning (RL) techniques which are used to attempt to overcome time degeneration of rule efficacy. RL amends programs incrementally in real time by rewarding success in the course of production execution. "Rewards" reinforce the probability of the repetition of a successful action on future occurrences of a given state. In this paper the RL is the well-tested Actor-Critic methodology. While online the GNP-RL combination makes dynamic choices as to the timing of trades and which combinations of technical statistics should dictate it.

Against 20 of Japan's largest stocks over 2001-2004 (again a volatile index with poor overall returns), with periods of 2.75 years training, 3 months validation and 12 months testing profits of 7.1% were gained against the buy-and-hold's 1.5%.

Again in 2007 [Chen,2007/2] experiment with GNP and a second recognised RL algorithm, State-Action-Reward-State-Action (Sarsa). This is compared with GNP on its own and with buy-and-hold. An intricate function is described for informing Sarsa, whilst online, through the dynamic evaluation of the relative suitability of many technical indices as criteria in each given instance. Sarsa dynamically adjusts the weighting of the multiple decision nodes and links evolved by GNP. Testing involves 16 of the 20 firms used previously over the same periods.

For most of the stocks (13/16) simple GNP outstrips buy-and-hold and the Sarsa addition improves markedly on this. In all 5 stocks where buy-and-hold loses money GNP-Sarsa takes profits. The most interesting evidence is that of steady increase of fitness and performance over the

generations after training, during production use.

[Yin,2007] experiment in generating call and put options on 2006 FTSE100 futures prices, with plain GP. However they address another fundamental drawback of GP: that for each focused application of GP the modellers work through trial and error to choose the most appropriate settings of the many parameter choices in the algorithm, but that the most productive choices may well be entirely different in different areas of the data investigated and different times in the run. Here a methodology is tested for the dynamically adaptive co-evolution of those parameters under influence of feedback from the evolution process.

Initialisation, selection and replacement strategies are the same for both comparison sets as are pop=300, gen=800, and cut=40, and the terminal and operator sets. But crossover and mutation rates are changed continually by feedback depending on measures of level of change and lack of new best rules over a given number of generations. (In general different levels of population and of problem complexity require different rates for these parameters.) Thus population search convergence at a threshold of a local optimum can be detected and escaped "whilst permitting local improvement around just discovered new solutions".

Further work suggested is in the dynamic co-evolution of the parameters (e.g. no change for ***n*** generations), for triggering the dynamic co-evolution of parameters!

Another feature used here is the integration of expert domain knowledge through the pre-calculation of various fixed specialist analytical measures into the terminal set.

Experimental results show appreciable improvement on performance gained by the new methodology over ordinary fixed GP.

Financial Text Mining and GP :

A review of several of the latest papers in this sub-field which has been chosen as the focus of this project, points up many of the variables, choices and difficulties to be addressed in the program design cycle for the project but which can only be made on the basis of practical possibilities revealed at the time of design.

[Fung 2003] identify a growing emergent field of research in stock price prediction through text mining, where evidence has been presented of a relationship between market movements and news articles, referencing among others his own slightly earlier work and [Thomas,2002].

As an advance on this he suggests the mining of interrelationships between news time series on multiple stocks or topics influencing the price of a single stock (rather than a simple 1 to 1 relationship). A simple example is patterns of news that are shown to trigger Oracle's price movements may be as much a predictor of Sybase movements as articles on Sybase.

There is interesting discussion in relation to EMH holding that the market will efficiently process and adjust for newly broadcast information immediately, whereas on a human level this plainly can't be so, leaving room for profitable machine trading.

TFIDF weighting and Support Vectors Machine (SVM) software is used for generating the relationships between news features and trends and signalling of either rises or falls in stock price (600,000 Reuters Market 300 Extra news articles over 6 months in 2003 versus 33 HangSeng stocks, with 1 further month testing). These are then analysed using time series segmentation to seek inter-relationships between stocks. Allocation of articles to topics, sectors and specific stocks is conveniently managed by thorough hypertext tagging by Reuters.

The comparisons are between the generated multiple and single stocks' time series predictions and also buy-and-hold. The respective average gains of 6.5%, 3.7% and -2.6% (falling market again), are encouraging for their methodology.

The strategy of buy for 24 hours then sell for a positive signal, and the reverse for a sell signal, is an interesting simplification that is not discussed though it points to the complexity of measuring the duration of an effect and its time-delay from the signal. Further work is suggested on what constitutes too few articles aligned to a stock for prediction, and too many (noise), and for relationships between and within sectors.

[Larkin, 2008] attempt to exploit the quantification of sentiment in real-time news data stream items, recently achieved by RavenPack International, S.L. [http://www.ravenpack.com/]. They choose Dow Jones Network news items, characterised as positive, negative or neutral toward a particular sector. GP is then applied to this data stream, with no other data input, to achieve predictive rules for immediate price reactions in the S&P500. Almost uniquely they try to exploit only intra-day movements. Their GP rules successfully predicted price jumps up to an hour before the market reacted to news items.

As preamble they point out that high quality information unknown to others, can of course not be subject to the fabled EMH ability to instantly reflect all information. There is a finite number of people with the ability to analyse even numerical information, but this must be much more the case where a human interpretation of text is required. Beating such diffusion is the job of all predictive systems. Surprisingly, it is pointed out, there is no simply visible correlation between news and market movements.

The quantity and intricacy of both data preprocessing/normalisation and fitness function decisions are a little troubling here, combined with the proprietary input. Data is from S&P500 2007 at one minute's resolution. GP inputs include news hits in average values per minute over the last 20 minutes, 60 minutes and 1 week. The fitness function is simple prediction of whether the price beats two standard deviations of the mean any minute in the next hour. (Pop=500, gen=51).

The base-case model for comparison is a distribution aware random predictor. Its positive calls were 1/24 correct against 1/3.6 for the GP (but in the 2.6 falses are many where the rise is 0.1 to 1.9 standard deviations). The results were found to be statistically significant at $P < 0.001$. Most interestingly, as so often, results would be very much better using only "volatile periods" (and the last 1.5 hours of the day!).Future work suggested is combining the usual price and volatility data feeds into the GP.

Whilst Larkin et al refer to several preceding EA stock prediction papers and to several analyses of the impact of news and commentary upon market movements from the purely financial domain, none is made to preceding work on computer text mining in direct relation to market reactions. However [Takahashi, 2007] study 13,000 Headline News on-stream broadcasts from JIJI PRESS and use automated text classification to analyse stock price reactions and claim successful use of the information for stock price prediction. [Rachlin, 2006] create a stock trading system using "prediction models...based on the content of time-stamped web [news] documents in addition to traditional Numerical Time Series Data ...using several known classification algorithms". Finally [Geller,2007] has done similar sentiment analysis and rating in relation to individual company names across the domain of personal internet blogs and forums. Although this latter comprises only a search engine for companies to gauge statistics and scores against their name for the type of sentiment expressed this opens possibilities for further data sources for the work of [Larkin,2008], assuming rigorous human selection of the blogs and forums used.

[Thomas,2002] extending upon their similar work in 2000 measure volume of postings on the stock specific message boards of the financial discussion areas of both yahoo.com and ragingbull.com, in order to predict stock movements on 68 stocks which are selected from the

Russell1000 as being those 10% with the highest message volume (1998-2001 inclusive).

Half of the companies are for training, half for testing. Validation, on 50% of the training data, is done on the whole population (just 20) at each generation (just10), taking the best rule only on the last generation. A maximum of 10 nodes are allowed.

The approach contrasts greatly with [Larkin,2008] in that it only counts posts made whilst markets are closed due to the ("Efficient Market"!) assumption that all information published during the trading day will be efficiently factored into the price immediately.

Returns are the only fitness measure and the only output is a binary decision to either long or short the stock. If message traffic volume in the period between closing and opening of the market is greater than a threshold the stock is shorted for a given number of days (a parameter fed into the algorithm).

The threshold levels are also parameters (n,m) fed into the GP : ***n*** standard deviations above the mean message count, over ***m*** days window to determine the mean and deviations.

Null Hypothesis/p-value testing is presented to show that the fairly impressive risk adjusted(Sharpe ratio) excess returns are statistically significant and thus the data has predictive power. A reasonable statistical argument is also presented to show this data contributes information not present in other numerical data, particularly price and volume for this data set.

As with many GP experiments the feeling is left that too little testing in too little data has been done to create general conclusions beyond special conditions in the period or stocks used. In this case almost all the stocks are tech stocks (hence the message volumes) and all of the excess returns in the test period are visibly inside the most extreme escalation in prices during 2000. Excess returns plunge at the time of the dotcom crash. As with so many models it appears that a model, and not just a best rule, needs to be selected for each set of market conditions.

Bibliography and References:

Search sources for materials have been:

- ISI Web of Knowledge database.
- EI Compendex Engineering Village database.
- ScienceDirect database.
- "The Genetic Programming Bibliography", a near comprehensive listing at http://www.cs.bham.ac.uk/~wbl/biblio/ actively maintained by **William B. Langdon.**
- Catalogue searches of Edinburgh and Heriot-Watt Universities.

[Allen,1999] ALLEN, F. & KARJALAINEN, R. (1999) Using genetic algorithms to find technical trading rules. *Journal of Financial Economics,* 51**,** 245-271.

[Becker,2003] BECKER, L. A., SESHADRI, M. (2003) GP-evolved technical trading rules can outperform buy and hold. *Proceedings of the 7th Joint Conference on Information Sciences***,** 1136-1139.

[Becker,2003/2] BECKER, L. A., SESHADRI, M. (2003) Comprehensibility & Overfitting Avoidance in Genetic Programming for Technical Trading Rules. Technical report, Worcester Polytechnic Institute, 2003.

[Chattoe,2004] CHATTOE, E. (2004), web review of CHEN, S. H. ed. (2002) *Genetic Algorithms and Genetic Programming in Computational Finance,* http://jasss.soc.surrey.ac.uk/7/4/reviews/chattoe.html , viewed 23rd May 2008.

[Chen,2003] CHEN, S. H. & KUO, T. W. (2003) Overfitting or poor learning: A critique of current financial applications of GP. *Genetic Programming, Proceedings,* 2610**,** 34-46.

[Chen,2007] CHEN, Y., MABU, S., HIRASAWA, K., HU, J., (2007) Genetic network programming with sarsa learning and its application to creating stock trading rules. *2007 IEEE Congress on Evolutionary Computation, CEC 2007*, *2007 IEEE Congress on Evolutionary Computation, CEC 2007*, 2008, 220-227.

[Chen,2007/2] CHEN, Y., MABU, S., HIRASAWA, K., HU, J., (2007) Stock trading rules using genetic network programming with actor-critic. *2007 IEEE Congress on Evolutionary Computation, CEC 2007*, *2007 IEEE Congress on Evolutionary Computation, CEC 2007*, 2008, 508-515.

[Dempsey,2002] DEMPSEY, I., O'NEILL, M., BRABAZON, A. (2002) Investigations into market index trading models using evolutionary automatic programming. *Artificial Intelligence and Cognitive Science, Proceedings,* 2464**,** 165-170.

[Dempsey,2006] DEMPSEY, I., O'NEILL, M., BRABAZON, A. (2006) Adaptive trading with grammatical evolution. *2006 Ieee Congress on Evolutionary Computation, Vols 1-6***,** 2572-2577.

[Fung,2003] FUNG, G. P. C., YU, J. X., LAM, W. & IEEE (2003) Stock prediction: Integrating text mining approach using real-time news. *2003 Ieee International Conference on Computational Intelligence for Financial Engineering, Proceedings***,** 395-402.

[Geller,2007] GELLER, J., PARIKH, S. & KRISHNAN, S. (2007) Blog mining for the fortune 500. *Machine Learning and Data Mining in Pattern Recognition, Proceedings,* 4571**,** 379-391.

[Larkin,2008] LARKIN, F. & RYAN, C. (2008) Good news: Using news feeds with genetic programming to predict stock prices. IN ONEILL, M., VANNESCHI, L., GUSTAFSON, S.,

ALCAZAR, A. I. E., DEFALCO, I., DELLACIOPPA, A. & TARANTINO, E. (Eds.) *Genetic Programming, Proceedings.*

[Pavlidis,2007] PAVLIDIS, N.G., PAVLIDIS, E.G., EPITROPAKIS, M.G., PLAGIANAKOS, V.P., VRAHATIS, M.N. (2007) Computational intelligence algorithms for risk-adjusted trading strategies.
Evolutionary Computation, 2007. CEC 2007. IEEE Congress on Evolutionary Computing, 540-547

[Potvin,2004] POTVIN, J. Y., SORIANO, P. & VALLEE, M. (2004) Generating trading rules on the stock markets with genetic programming. *Computers & Operations Research,* 31**,** 1033-1047.

[Rachlin,2006] RACHLIN, G. & LAST, M. (2006) Predicting stock trends with time series Data Mining and Web Content Mining. *Advances in Web Intelligence and Data Mining,* 23**,** 181-190.

[Takahashi,2007] TAKAHASHI, S., TAKAHASHI, M., TAKAHASHI, H. & TSUDA, K. (2007) Analysis of the relation between stock price returns and Headline News using Text Categorization. *Knowledge-Based Intelligent Information and Engineering Systems: Kes 2007 - Wirn 2007, Pt Ii, Proceedings,* 4693**,** 1339-1345.

[Thomas,1999] THOMAS, J., SYCARA, K., (1999) The importance of simplicity and validation in genetic programming for data mining in financial data. IN FREITAS, A. (ed) Data Mining with Evolutionary Algorithms: Research Directions, pages 7-11, Orlando, Florida,Technical Report WS-99-06.

[Thomas,2002] THOMAS, J., SYCARA, K., (2002) GP and the Predictive Power of Internet Message Traffic. IN CHEN, S., (ed) Genetic Algorithms and Genetic Programming in Computational Finance, chapter 4, pages 81-102. Kluwer Academic Press.

[Yin,2007] YIN, Z. **,** BRABAZON, A., O'SULLIVAN, C. (2007) Adaptive genetic programming for option pricing. *Proceedings of GECCO 2007: Genetic and Evolutionary Computation Conference, Companion Material*, *Proceedings of GECCO 2007: Genetic and Evolutionary Computation Conference, Companion Material*, 2007, p 2588-2594

==========

5) Statement on professional, ethical and social issues.

The project involves the development of well known Genetic Programming algorithms. Basic implementations of these algorithms in Java code are available on the internet and copies of these will be used to achieve simple working starter programs. All code used will be checked for commercial restrictions, fully attributed to source in both the the code and the dissertation, and clearly differentiated from additions and amendments for the project.

The project does not require any personal data to be stored, gathered or accessed. All of the data being used is openly and permanently available in the public domain on the Web. The text of publicly available news, comment and blog items is only used for statistical analysis and not reproduced in any way.

There will be no other people involved in the project in any role other than the author and supervisor and the only equipment used will be personally owned PCs at home.

The social and ethical as well as legal issues would need to be reconsidered in detail in the unlikely case of such successful results as would entail the real-world commercial use of the algorithms!

6) Conclusion and issues.

The project's aims, objectives, rationale and structure have been set out in summary in this portfolio document as a starting point for work on software design and implementation of the statistical experiment. A broad indication of the content and structure aspired to for the evaluation of these, and for the dissertation has been given. A review of background literature has framed the knowledge base and issues around which the software will be built, though the principle selective constraint in this must be development time available, which will become clearer in the next detailed planning phase of the project.

This project is entirely software based, therefore all risks in the project are related to either time, or the successful obtaining of the required software, publicly available development training, and of adequate data. In particular a supply, or several, of news or comment of adequate quality, length of history, and if possible, proprietary preprocessing, needs to be sourced at a reasonable cost as a first step in detailed planning.

The attempt to combine statistics relating to meaningful web documents or text with the widely available organised numeric data that is commonly used in financial GP applications is possibly an ambitious target for development and analysis within the 460 hours available.

In the event that any of these risks have a negative outcome during the detailed planning stage, the contingency plan is to pursue the objective of developing a simpler predictive financial GP application. This will attempt to combine a small selection of the innovations and approaches recently suggested, to attempt to overcome some of the shortcomings and poor results of GP in the field, as described in the literature review. The results of this against a new set of financial data will be analytically compared to results from other machine learning techniques in order to assess their level of statistical significance, as suggested by [Chattoe,2004], and Corne(supervisor).

==========

7) Risk assessment form.

See Reverse.