Development of 2D Curve-Fitting Genetic/Gene-Expression Programming Technique for Efficient Time-series Financial Forecasting

Manal Alghieth
Faculty of Technology
De Montfort University
Leicester, United Kingdom

Yingjie Yang
Faculty of Technology
De Montfort University
Leicester, United Kingdom

Francisco Chiclana
Faculty of Technology
De Montfort University
Leicester, United Kingdom

Abstract-- Stock market prediction is of immense interest to trading companies and buyers due to high profit margins. Therefore, precise prediction of the measure of increase or decrease of stock prices also plays an important role in buying/selling activities. This research presents a specialised extension to the genetic algorithms (GA) known as the genetic programming (GP) and gene expression programming (GEP) to explore and investigate the outcome of the GEP criteria on the stock market price prediction. The research presented in this paper aims at the modelling and prediction of short-to-medium term stock value fluctuations in the market via genetically tuned stock market parameters. The technique uses hierarchically defined GP and GEP techniques to tune algebraic functions representing the fittest equation for stock market activities. The proposed methodology is evaluated against five well-known stock market companies with each having its own trading circumstances during the past 20+ years. The proposed GEP/GP methodologies were evaluated based on variable window/population sizes, selection methods, and Elitism, Rank and Roulette selection methods. The Elitism-based approach showed promising results with a low error-rate in the resultant pattern matching with an overall accuracy of 93.46% for short-term 5-day and 92.105 for medium-term 56-day trading periods.

Keywords: genetic programming; gene expression programming; Time series Stock market prediction; stock market.

1. Introduction

Stock market predictions have been of high interest to traders due to the potential of gaining high returns in relatively short periods of time. Stock market predictions are one of the most frequently attributed factors used in the financial markets by companies and individuals to engage in profitable stock trading. Rise and fall of stock prices is not regarded as a linear system and is attributed to a large and diverse range of factors. This makes a direct or expert-driven rule-based approach to stock trading an extremely tedious and unreliable technique.

There have recently been major studies in the domain of artificial intelligence and pattern recognition for the prediction of time-series-based financial activity patterns in the stock market. The majority of these efforts have mainly focused on artificial neural networks (ANN), hidden Markov models (HMM), and fuzzy inference/logic (FL) techniques. Moreover, areas of statistical analysis including temporal trend analysis via support vector machines, curve fitting via Kalman filters and graphical splines have already been explored and offer promising outcomes (Bisoi *et al.*, 2011).

The evolutionary computing domain has also been extensively exploited for the stock market prediction algorithms. However, general issues of over-fitting, data snooping bias and black-boxed characteristics are cited as common challenges regarding the reliability of these algorithms. Genetic Programming (GP) has been used to generate trading rules as an alternative to the well-known buy-and-hold approach of stock buying. Potvin *et al*. (2004) focused on individual stock-based adjustment technique to predict short-term fluctuations. Long-term prediction, on the other hand, offers its own complexities due to over-time error accumulation and inherent uncertainty due to large time-durations involved. Therefore, in stock prediction, long-term prediction is taken in a completely different perspective compared to the medium-to-short term case. Sovilj *et al*. (2010) utilised optimally pruned extreme learning machine (OPELM) and optimally k-nearest-neighbours (OPKNN) models that are known for their fast training times and accurate predictions. The former OPELM algorithm relies on a single-layer fast feed-forward neural network, whereas the later OPKNN algorithm relies on k-nearest neighbours as the kernel function. OPKNN algorithm is faster than its former counterpart OPELM algorithm due to its deterministic nature as it does not use any additional parameters. Another approach to long-term time-series forecasting was proposed by Stojanović *et al*. (2014) based on mutual information (MI) used in regression tasks for feature selection. The main contribution of this approach has been to differentiate useful data from outliers by recursively reducing uncertainty in the training set by eliminating noisy datasets from the training stage.

In the domain of stock forecasting, Gene Expression Programming (GEP) is addressed in the existing literature by Bautu *et al.* (2010), Garg *et al.* (2013), Ye *et al.* (2009) and Hongbin *et al.* (2010). However, the majority of these works focus on optimising ensembles of other AI algorithms such ANN, Support Vector Machines (SVM) and Naive Bayes.

Bautu *et al.* (2010) primary focused on modelling an ensemble of classifiers based on the Efficient Market Hypothesis (EMH). The theory is based upon the fact that it is not possible to beat the market because the stock market efficiency always incorporates all the factors affecting the market such as socio-political changes,

natural disasters and other unexpected localised fluctuations. This means the stocks always trade at their fair value, which makes it impossible for traders to purchase shares at lower-than-market or higher prices. Consequently, it is not possible to outperform market via expert-driven stock selection or market timing. Therefore, EMH states that the only way to obtain better returns is by purchasing higher-risk investments. The underlying idea of this work can lead to generating best performance equations from training data that can then be used as relative fitness functions in the proposed work.

Bautu *et al.* (2010) utilised an evolutionary approach to uncover quantitative patterns in the stock market performance. Contrary to the proposed approach, AdaGEP approach in this work utilises logical expression trees with a binary classification approach. The approach achieves diversity by starting the genetic run at random seeds. The approach uses an ensemble of GEP classifiers that run against input/output data pairs. This approach, despite its merits of an increased confidence in multiple-models running in parallel increasing the classification reliability, is still time-consuming compared to an approach where an optimal solution is obtained via single genetic classification run operating on a generation of chromosomes that started at their own pace. Hence, GEP in this approach is not directly used to predict certain stock prices but to identify the most optimal set of other classifiers including Naive Bayes, SVM, and Multi-layer Perceptron, that provided the best classification outcome.

Similarly, extended work done by Barbulescu and Bautu (2012) also focuses on combining ARMA and GEP-induced models to exploit ARMA's capability to identify linear trends and GP to classify nonlinear trends in the data. As the focus of this research primarily stays on ARMA models, the identification and estimation is highly likely to be distorted by outliers. Additionally, work done by Grosan and Abraham (2006) also proposes the performance analysis of an ensemble of four measures namely Root Mean Squared Error (RMSE), Maximum Absolute Percentage Error (MAP), Correlation Coefficient (CC), and Mean Absolute Percentage Error (MAPE). Similarly, the work by Garg *et al.* (2013) also used GP focussing on model selection criterion and data transformation. Hongbin *et al.* (2010) focussed on a hybrid GEP-ANN approach with an objective to improve complex problems, enhance learning speed and generalise of Back Propagation ANN. The ultimate objective of this research was to improve the prediction accuracy of ANN.

Unlike the GA and GP work to-date, the proposed method here aims at a more direct approach to solving stock value prediction by using genomes whose strings of numbers represent symbols. These strings of symbols can further be expressed as equations, grammars or local mappings. The genome is then mapped to a binary tree which can be walked-through to evaluate the underlying equation.

Thus, contrary to the GP and GEP techniques in the literature, the technique put forward here builds a model predicting the future function values based on the previous history of stock values. The technique calculates closed-form equations of the most optimal dataset from real-world stock market data. Therefore, the closed-form equations obtained are the comparative fitness-functions for the GEP run. The underlying principle is extremely powerful as, unlike conventional GAs, it does not map to hard values but to symbols that iteratively combine to give equations that are nearest to the closed-loop fitness equations derived from the existing training data. In order to develop closed-form equation stock value variations were treated in two different modes namely the long-term and the short-term mode.

2. Basic Concepts

The main stock prediction problem can be described as follows: Provided current stock value(s) and the trading volume, the task is to predict the next time-instance stock market value. The hypothesis assumed here is that the stock values during a trading day change gradually and display an organised relationship between current and future stock values. This relationship can be described as an algebraic function whose derivation is the main objective (Peng *et al.*, 2014).

In GEP, two algebraic functions representing linear and exponential equations are used to derive linearly and exponentially regressive equations. These equations represent training data as lines on a 2-D plane that are used to represent a relationship between the current and future stock value (RegCal 2015).

Day	1	2	3	4	5	6	7	8
Closing Value	21.11	22.05	21.56	21.49	22.31	22.78	23.17	23.76

Figure 1: Representation of 7-day time-series data

For instance, for data shown in Figure 1, mapping the days on the x-axis and Day Closing Value on the y-axis, the equation for the curve can be obtained by the method of least square regression where the best fit line is computed via the equation given below:

$$y = m\,x + n \qquad (1)$$

Equation (1) satisfies the condition that the sum of the squared vertical distance between two points and the line is minimised. An exponential regression curve of this curve or the logarithmic regression curve is then obtained via the equation (2):

$$y = a + c \ln(x) \qquad (2)$$

The slope of the regression curve is thus given by the formula:

$$m = \Big/_{\sum x^2 - n(\bar{x})^2}^{\sum xy - n(\bar{x})(\bar{y})} \qquad (3)$$

Based on the equation (3), a correlation coefficient showing how close the line fits can be computed via the following equation:

$$\Big/_{\sqrt{(\sum x^2 - n(\bar{x})^2)(\sum y^2 - n(\bar{y})^2)}}^{\sum xy - n(\bar{x})(\bar{y})} \qquad (4)$$

The correlation coefficient value ranges from -1 to +1 where, if the correlation is close to zero, it means the data does not exhibit a linear relation. Similarly, equations for exponential, power and logarithmic regression curves can

be transformed in their respective linear forms. For instance, equation can be linearly transformed by taking the natural logarithm of both the sides.

Hence, for data shown in Figure 1, linear and exponential equations can be obtained as follows,

$$Linear: y = 0.2996 * x + 20.8686 \quad (5)$$
$$Correlation: 0.8741$$

$$Exponential: y = 20.8951 * 1.0136^{x} \quad (6)$$
$$Correlation: 0.8731$$

Based on equations (5) and (6), the data is mapped in input/output pairs and the following symbols shown in Figure 2:

	1	2	3	4	5	6	7	8
Symbols	a	b	*	/	+	-	exp	sqrt

Figure 2: Representation of genome symbols for the proposed GEP approach

Based on the abovementioned symbols in Figure 2, the GA attempts to derive the best set of symbols for a genome with the highest fitness. The fitness function in the run aims to generate the highest number of results that occur while traversing the gene tree and generating values that are closest to the correlation generated by the equations shown in (5) and (6). Hence, the fitness function in this case is the summation of correlation difference and subtraction it with a large number as follows:

$$fitness = \sum_{i=1}^{100} abs[(corr_{(5)} + corr_{(6)}) - (corr_{(obs-5)} + corr_{(obs-6)})] \quad (7)$$

Therefore the minimum the difference of correlation between a gene-built equation and the equations (5) and (6) the higher the fitness of the respective solution will be. According to RegCal (2015), two additional equations based on power and logarithmic equations can also be evaluated against a GEP run.

The various techniques/models used to predict time-series data in the proposed technique are explained as follows. The selection method aims to keep a single or a set of solutions in the generation while moving from one to another. Elite, Rank and Roulette wheel selections are three selection methods, with the justification of Elite already explained above. Rank-based selection extends the main Roulette Wheel selection methodology. The Roulette Wheel selection mainly works as a fitness proportionate selection. However, this results in additional bias toward fittest individuals which reduces the overall diversity of the population. Rank-based selection uses fitness-based ranking which provides ranks regardless of how "least-fit" an individual is. The function sets used in the methodology were as shown in Figure 3.

	1	2	3	4	5	6
Simple	a	b	*	/	+	-

	1	2	3	4	5	6	7	8
Extended	a	b	*	/	+	-	exp	sqrt

Figure 3: Simple and extended function sets used in the GEP

Based on the capability of GEP to evolve and improve programmatic expressions, the rationale behind the usage of GEP for stock prediction can be elaborated into three points:

1. Most importantly, the chromosomes representing solutions are simple entities which can be replicated (i.e. recombined, transposed, mutated) to generate probabilistically more efficient solutions. This approach offers a promising potential to predict forecasting based on time-series data based on equations trained on existing training data.
2. The expression trees (ET) formed in GEP express chromosomes which then change based on crossover. This "tree-walking" operation reproduces genetically evolving equations which can be compared to original curve-fitting equations to estimate fitness values of each chromosome. Again, contrary to standard brute-force search mechanisms, this ability can iteratively estimate iterative fitness of all the chromosomes in a generation. The trait makes it possible to generate multiple best solutions that can then be used to predict a forecast value based upon previous data.
3. According to Ferreira (2001), the GEP algorithm surpasses GP by more than four orders of magnitude. The performance improvement of GEP is of crucial importance particularly during the processing of massive real-world datasets.

3 Evolutionary prediction of stock data via GA, GP and GEP

GAs operate on the gradual improvement of a set of solutions in a user-defined fitness hill-climbing process. The algorithms were originally introduced to address NP-Completeness in algorithms to which a direct solution via brute-force is not possible or extremely difficult in polynomial time. A few well-known algorithms are travelling sales person, knapsack fitting and location-allocation problems. A generic stock exchange value prediction, based upon n number of stock parameters and m number of features, result in a computationally complex problem for which an optimal is not possible in polynomial time.

3.1 Shortcomings of evolutionary hybrid methodologies in time-series prediction

Despite having a large body of literature covering GAs, reliability of GAs in future time series predictions is still looked upon with doubt. GAs can often present over-fitted results based on existing temporal data. The literature does show an increased application of evolutionary computing techniques to hedge funds. However, based on safety issues, the models are extensively evaluated against back-tests before they are actually used in actual production cases. This often leads to months of testing before a model is actually allowed to run on real-life cases. Model reliability in similar situations is improved by separating the sampling universes while keeping the confirmation back-tests separate. The strategy is to make a GA from data till time t and then subsequently testing it till $t + N$ before actually trusting the underlying model.

As discussed earlier, despite its own strength, the three biggest weakness of the GA are: data over-fitting, data snooping bias and black-box operation.

3.1.1Data over-fitting

Similar to conventional ANNs, data over-fitting in GEP occurs when possible crossovers between a fitter and weaker chromosome take place resulting in gene parameters generating weaker solutions in the subsequent generations. This can be controlled via a number of ways. The simplest form is to monitor a genetic run for its learning error and stop the run as soon as a consecutive rise in error rate is noted for a number of runs.

However, data over-fitting can be avoided in three ways: 1) Providing an indirect bias towards simplicity of solutions or setting a penalty against complex solutions, 2) Restricting the number of models considered in a run, and/or 3) Using a validation dataset. The lateral validation dataset cannot be an ideal case for a time-series case, as a validation dataset may belong to a different time with variable circumstances which do not reflect the overall company stock profile (Chan, 2011).

3.1.2 Data snooping

The problem of data snooping in GA can be avoided by well-organised mutation operators/functions. However, setting a specific mutation rate to avoid snooping is another challenge where a very high mutation rate may reduce a GA run into a randomised search. On the other hand, if the mutation probability is too low, the function does not serve its purpose of diversifying the genetic population.

3.1.3 Black-box operation

An average evolutionary run acts as a black-box where a continual hill-climbing process finally terminates into an optimised fitness function. A GA run cannot be changed apart from its specialised selection operators (e.g. crossover and mutation) in order to prevent the algorithm from either turning into a random search algorithm or getting stuck into a local minima.

3.2 GA-based GP/GEP Architecture for Time Series Prediction

Based on the description of the problem provided in the above sections, the adaptation of GP/GEP to the proposed algorithm is given below:

3.2.1 Chromosomal Encoding

A set of functions F made of classical algebraic, Boolean and relational programming parameters including logical decision constructs that include IF-THEN-ELSE and a range of programming functions are used in the encoding stage. For a set of terminals $\boldsymbol{\tau}$ related to various constant types related to stock opening, high, closing, and trading volumes are described as follows:

- Arithmetic operators: $\{+, -, *, /\}$
- Boolean operators: $\{\&\&, ||, !\}$ representing AND, OR and NOT respectively
- Relational operators: $\{<, >\}$
- Boolean functions: $\{if - then - else\}$
- Real functions: (See below)
    - $normal(r_1, r_2\}$: Price difference between two real numbers
    - $average(share - price, days)$: Average stock price over the past n days
    - $maximum(max - price, days)$: Maximum stock price over the past n days
    - $\mathrm{minimum}(\text{min-price}, \text{days})$: Minimum stock price over the past n days
    - $delay(max - price, days)$: Closing stock price delayed by n days
    - $RSI(days)$: The relative strength indicator during the past n days as given by Potvin *et al*. (2004)
    - $ROC(days)$: Rate of change of stock price as given by Potvin *et al*. (2004)

3.2.2 Fitness Evaluation

The most basic fitness rule in this methodology is taken to be the rate of change of stock prices over a certain historical data pattern (medium or short-term).

The fitness criterion in this research mainly focuses on calculating the inverse of linear and exponential correlation of genetic equations created as a result of symbolic tree walk to the actual equations generated/created via the 2D stock data.

3.2.3 Generation of initial population

The methodology uses an ensemble of programming trees that follow the recursive construction processing from root to leaf via the guidelines described by Potvin *et al*. (2004).

- The tree root selection is made from Boolean functions and operators.
- After the root has been selected, its descendants (leaves) can be selected from Boolean constants and functions and Boolean or relational operators.
- If a relational operator is selected, its descendants are selected from either real functions or terminals.

3.3 Selection of next generation chromosomes

The selection process in this work adopts three main methods namely elitism, rank-based selection and the roulette-wheel sampling (Whitley, 1989; Baker, 1985).

In rank-based sampling, during any genetic run, all the programmes (chromosomes/solutions) are ranked based upon their best to worst (p) raw fitness values. For each chromosome, a new fitness value is then assigned based upon the following formula:

$$F_i = Max - [(Max - Min)(i - 1)/(p - 1)] \quad (8)$$

In roulette-wheel selection, each programme is assigned a slice of the roulette wheel based upon its fitness. Therefore, the individuals with highest fitness have more chances of getting selected in the subsequent generations.

Finally, the elitism-based methodology is used to select at-least one best individual with the highest fitness value in

the subsequent generations. This minimises the chances of best programmes recombining with weaker candidate resulting in best solutions getting lost during the genetic runs.

4 Results and Discussions

As elaborated earlier, the main focus of this research was on the critical analysis of GEP sliding window algorithm against a diverse range of datasets and varying prediction scenarios ranging from input window containing 2+ month trailing stock data to short-term data looking back into just a week in past. The data was selected for five well-known stock companies Yahoo, British Petroleum, Glaxo Smith Klien (GSK), HSBC and the RBS. In order to understand the difference of these companies' performance, it is necessary to look into the nature of performance they showed during the past 20+ years. The GEP prediction model for Yahoo compared with the previously reported neuro-fuzzy model. All the dataset described below were trained and evaluated over two window sizes of 5 days (short-term) and 56 days (medium-term).

Figure 4 represents stock data from Yahoo and British Petroleum where the later at present shows a closing price of 523.9, a dropping trend of 6.28 at 1.2%.

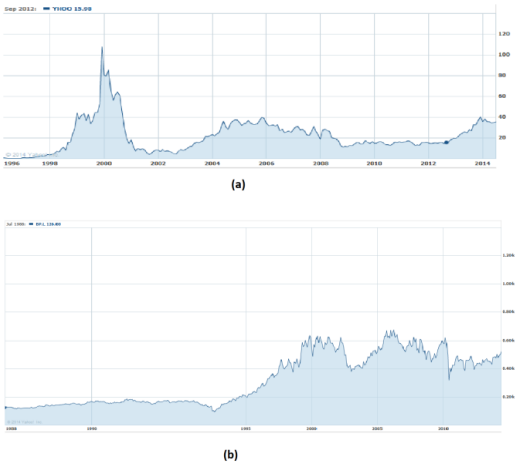

Figure 4: Yahoo and British Petroleum stock data

GSK was selected on its unusually spontaneous stock turmoil during early 1990s due to increasing competitions from Wellcome PLC which launched a large number of new pharmaceutical drugs during 1980s and early 1990s. The stock currently stand at a falling pattern of 14 (0.89%) at the stock price of 1562.95 (Figure 5).

Figure 5: GSK stock data

The two banks involved in these tests were HSBC and RBS. Data showed a phenomenally large drop in shares during the past 5 recession years (See Figure 6). The shares continue to fall at a respective prices of 11.02 (1.96% fall) and 604.5 (0.15%) for RBS and HSBC PLCs.

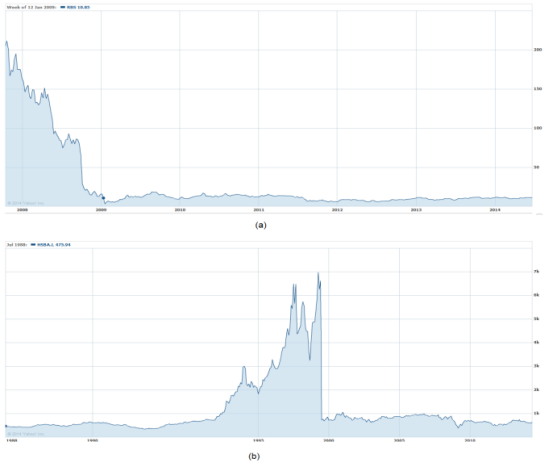

Figure 6: HSBC/RBS stock

In the above four cases, it must be noted that both medium-term and short-term predictions bear important insight for traders to buy or sell certain shares. For instance, for long-term investors, an accurate and reliable prediction algorithm representing a correct trend (increasing or decreasing) will support in the decision making process.

The performance of the proposed sliding window GEP algorithm was evaluated under the following conditions:

- Ability to predict existing pattern (increasing/decreasing) based on
  - Short-term training
  - Medium-term training
- Ability to predict the next-day stock prices as accurately as possible

4.1 Initial Selection Operator Evaluation

The algorithm's performance was initially evaluated over a fixed number of chromosomes with a total of 50 individuals (chromosomes) per generation and 1000 genetic runs with each performing/applying genetic crossover and mutation to induce diversity in the subsequent generation. The selection criteria were initially based upon elitism, rank-based and roulette-wheel algorithms. As shown Table 1, apart from the RBS case, the elitism-based selection performed better with the lowest learning errors for all the cases. Prediction errors were not focused at this stage because of possible improvement later-on once the remaining parametric combination (i.e. window-size, chromosome/generation, etc.) were evaluated.

Table1: Learning and prediction error rate for three selection criteria of Elitism, Rank-based and Roulette-Wheel selection

Company	BP			GSK			HSBC/A			RBS			Yahoo		
Chromosomes /generation	50			50			50			50			50		
Window Size/Category	5			5			5			5			5		
Genetic iterations	1000			1000			1000			1000			1000		
Selection Method	Elite	Rank	RW	Elite	Rank	RW	Elite	Rank	RW	Elite	Rank	RW	Elite	Rank	RW
Learning Error	8.69	14.87	18.89	29.71	37.09	29.74	53.48	120	53.45	25.72	23	28.89	4.25	1.63	1.632
Prediction Error	4	13	11	48	40	46.99	17.03	0	17.99	11	8	9	2	1	1
Legend: RW: Roulette-wheel Note: Prediction accuracies rounded to the nearest whole number and represented in US$															

4.2 Evaluation over short and medium-term data spans (day/week durations)

Based on the abovementioned performance, the five datasets were tested against training performed under two window sizes of short 5-days and 56-day-lengths. Both combinations were repeatedly evaluated against the following cases:

- Simple and extended function sets
- Genetic programming and GEP

Table 2: Short-ahead next-day-prediction based on medium-term and short-term look-back periods

	Short-ahead prediction: Prediction of the next day					
	Gene Function/Arithmetic Operations	BP	GSK	HSBA	RBS	Yahoo
Medium-term (56 days)	Simple	11	40	0	3	4
	Extended	4	24	9	8	1
Short-term	Simple	0	48	18	3	3
	Extended	13	23	0	8	2

Table 2 presents a mixed response to the change of gene functions where simple arithmetic function and arguments do not seem to have an overall impact on the prediction accuracy. It must be noted that the values shown in Table 2 are the difference of stock value by which a prediction has given an erroneous value. The extended gene function class represents arithmetic functions $\mathcal{F} = \{\Re, *, /, -, +\}$ and some common arithmetic functions inclusive of the set $\{sin, cos, ln, exp, sqrt\}$. As discussed before, this methodology is used by the chromosomes to build arbitrary expression via genetic operators.

A critical view of the table, however, illustrates the fact that the extended functions do not substantially improve the overall prediction accuracy of stock trading systems at both the medium-term and short-term scales.

Considering the case of next week's prediction also predominantly confirms the validity of simple gene function apart from the case of GSK where the extended showed a substantial improvement in the error rates for all the cases.

Table 3: Medium-ahead 7-day-prediction based on long-term and short-term look-back periods

	Long-ahead prediction: Prediction of the week					
	Gene Function/Arithmetic Operations	BP	GSK	HSBA	RBS	Yahoo
Medium term (56 days)	Simple	4	60	9	8	2
	Extended	8	48	23	8	1
Short-term	Simple	8	26	0	3	1
	Extended	11	48	9	8	1

Table 4: Short and medium-term prediction accuracy based on GEP sliding-window operation (for 5-day look-ahead only)

	MT		ST	
	GP	GEP	GP	GEP
BP	89.39	87.01	96.23	89.76
GSK	76.34	100	76.65	100
HSBA	82.09	91.87	82.1	84.09
RBS	97.12	89.54	95	100
Yahoo	98.75	96.75	97.64	96.11
Overall accuracy	88.738	93.034	89.524	93.992

Table 4 finally shows a comparison between both the GA variants: the conventional GP and the GEP based classification. The comparison shows GEP to perform substantially better for both the cases of short and medium-term prediction.

5 Conclusion

The research presents an evolutionary methodology for the prediction of stock exchange data via a specialised extension to the conventional evolutionary GA. The approach was adopted to address well-reported problems of over-fitting, algorithmic black-boxing, and data-snooping issues via GP and GEP algorithms. Another issue analysed in this case was the suitability of the algorithms to predict daily and weekly stock market patterns based upon medium-to-short-term stock history. The outcomes presented an outstanding GEP accuracy compared to GP in stock prediction via simple, non-arithmetic algebraic expressions with the best performance reported at short-term forecasting of 93.992%. On the other side, the worst performance of 88.23% was reported on a GP algorithm at medium-term prediction which could be attributed to the fact the at 65-day-training-lengths, the algorithms routinely found data subjected to non-deterministic human-level changes due to global financial uncertainties such as the Middle East crisis and the global recession. However, it is understood that GEP can particularly be used to increase the overall confidence level of trading markets if it is properly evaluated for the period of a few months before actually being implemented in risky stock markets for behaviour prediction.

6 References

[1] Aladag C. H., Yolcu U., Egrioglu E., Bas E. (2014) Fuzzy lagged variable selection in fuzzy time series with genetic algorithms, Applied Soft Computing, Available online 29 April 2014, ISSN 1568-4946,

[2] Araújo R. A., Ferreira T. A. E. (2009) An intelligent hybrid morphological-rank-linear method for financial time series prediction, Neurocomputing, Volume 72, Issues 10–12, June 2009, Pages 2507-2524, ISSN 0925-2312.

[3] Baker, J. E. (1985) Adaptive selection methods for genetic algorithms In: Proceedings of the First International Conference on Genetic Algorithms and their Applications. Hillsdale, NJ: Lawrence Erlbaum, 1985. p. 101–111.

[4] Barbulescu A., Bautu E. (2012) A Hybrid Approach for Modeling Financial Time Series, The International Arab Journal of Information Technology, Vol. 9, No. 4, July 2012

[5] Bautu, E., Bautu, A., Luchian H. (2010) Evolving Gene Expression Programming Classifiers for Ensemble Prediction of Movements on the Stock Market

[6] Bisoi, R., Dash, P. K., Padhee, V., Naeem, M. H. (2011) "Mining of electricity prices in energy markets using a computationally efficient neural network," *Energy, Automation, and Signal (ICEAS), 2011 International Conference on* , vol., no., pp.1,5, 28-30 Dec. 2011 doi: 10.1109/ICEAS.2011.6147178

[7] Blandis E., Simutis R. (2002) Using Principal Component Analysis and Neural Network for Forecasting of Stock Market Index. Bizinesa augstskola Turiba SIA, Riga, 2002, 31-35.

[8] BP (2014) British Petroleum Plc [Online] Available at <https://uk.finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=3&f=2014&g=d&s=BP&ql=1> [Accessed: 20/06/2014]

[9] Cai Q., Zhang D., Wu B., Leung S. C. H. (2013) A Novel Stock Forecasting Model based on Fuzzy Time Series and Genetic Algorithm, Procedia Computer Science, Volume 18, 2013, Pages 1155-1162, ISSN 1877-0509,

[10] Cheng C., Chen T., Wei L. (2010) A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting, Information Sciences, Volume 180, Issue 9, 1 May 2010, Pages 1610-1629, ISSN 0020-0255,

[11] Cheng M., Andreas F.V. R. (2011) Evolutionary fuzzy decision model for cash flow prediction using time-dependent support vector machines, International Journal of Project Management, Volume 29, Issue 1, January 2011, Pages 56-65, ISSN 0263-7863,

[12] Ferreira C. (2001) Gene Expression Programming: A New Adaptive Algorithm for Solving Problems, Complex Systems, Vol. 13 Issue 2, p. 87-89.

[13] Garg, A., Sriram S., Tai, K. (2013) Empirical Analysis of Model Selection Criteria for Genetic Programming in Modeling of Time Series System. IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), p. 90-94.

[14] Gordini N. (2014) A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy, Expert Systems with Applications, Volume 41, Issue 14, 15 October 2014, Pages 6433-6445, ISSN 0957-4174,

[15] Grosan C., Abraham A. (2006) Stock Market Modeling Using Genetic Programming Ensembles, Genetic Systems Programming: Theory and Experiences, pp. 131-146

[16] GSK (2014) Glaxo Smith Klien Plc [Online] Available at < https://uk.finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=3&f=2014&g=d&s=gsk&ql=1> [Accessed: 20/06/2014]

[17] Hongbin, W., Liyi, Z., Haukui, W. (2010) The Research on Neural Network Prediction based on the GEP, Second International Workshop on Education Technology and Computer Science, p.362-365.

[18] Hong W., Dong Y., Chen L., Wei S. (2011) SVR with hybrid chaotic genetic algorithms for tourism demand forecasting, Applied Soft Computing, Volume 11, Issue 2, March 2011, Pages 1881-1890, ISSN 1568-4946,

[19] HSBA (2014) HSBC Holdings Plc [Online] Available at < https://uk.finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=3&f=2014&g=d&s=hsba&ql=1> [Accessed: 12/06/2014]

[20] Kim K., Han I. (2000) Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, Expert Systems with Applications, Volume 19, Issue 2, August 2000, Pages 125-132, ISSN 0957-4174,

[21] Kocadağlı O., Aşıkgil B. (2014) Nonlinear time series forecasting with Bayesian neural networks, Expert Systems with Applications, Volume 41, Issue 15, 1 November 2014, Pages 6596-6610, ISSN 0957-4174,

[22] Ni H., Wang Y. (2013) Stock index tracking by Pareto efficient genetic algorithm, Applied Soft Computing, Volume 13, Issue 12, December 2013, Pages 4519-4535, ISSN 1568-4946,

[23] Potvin J., Soriano P., Vallée M. (2004) Generating trading rules on the stock markets with genetic programming, Computers & Operations Research, Volume 31, Issue 7, June 2004, Pages 1033-1047, ISSN 0305-0548

[24] Pulido M., Melin P., Castillo O. (2014) Particle swarm optimization of ensemble neural networks with fuzzy aggregation for time series prediction of the Mexican Stock Exchange, Information Sciences, Volume 280, 1 October 2014, Pages 188-204, ISSN 0020-0255,

[25] RBS (2014) Royal Bank of Scotland [Online] Available at < https://uk.finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=3&f=2014&g=d&s=rbs&ql=1> [Accessed: 19/06/2014]

[26] RegCal (2015) Regression Calculation, [Online] Article available online at

<http://www.had2know.com/academics/regression-calculator-statistics-best-fit.html> Accessed 01/05/2015

[27] Smith C., Jin Y. (2014) Evolutionary Multi-Objective Generation of Recurrent Neural Network Ensembles for Time Series Prediction, Neurocomputing, Available online 12 June 2014, ISSN 0925-2312, http://dx.doi.org/10.1016/j.neucom.2014.05.062.

[28] Sourirajan K., Ozsen L., Uzsoy U. (2009), A genetic algorithm for a single product network design model with lead time and safety stock considerations, European Journal of Operational Research, Volume 197, Issue 2, 1 September 2009, Pages 599-608, ISSN 0377-2217,

[29] Sovilj D., Sorjamaa A., Yu Q., Miche Y., Séverin E. (2010) OPELM and OPKNN in long-term prediction of time series using projected input data, Neurocomputing, Volume 73, Issues 10–12, June 2010, Pages 1976-1986, ISSN 0925-2312,

[30] Stojanović M. B., Božić M. M., Stanković M. M., Stajić Z. P. (2014) A methodology for training set instance selection using mutual information in time series prediction, Neurocomputing, Volume 141, 2 October 2014, Pages 236-245, ISSN 0925-2312,

[31] Straßburg J., Gonzàlez-Martel C., Alexandrov V. (2012) Parallel genetic algorithms for stock market trading rules, Procedia Computer Science, Volume 9, 2012, Pages 1306-1313, ISSN 1877-0509,

[32] Venkadesh S., Hoogenboom G., Walter Potter, Ronald McClendon, (2013) A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks, Applied Soft Computing, Volume 13, Issue 5, May 2013, Pages 2253-2260, ISSN 1568-4946,

[33] Wei L. (2013) A GA-weighted ANFIS model based on multiple stock market volatility causality for TAIEX forecasting, Applied Soft Computing, Volume 13, Issue 2, February 2013, Pages 911-920, ISSN 1568-4946,

[34] Whitley D. (1989) The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best. In: Proceedings of the Third International Conference on Genetic Algorithms. San Mateo, CA: Morgan Kaufmann, 1989. p. 116–121.

[35] Xiao-qin W. (2012) Research on Prediction Model of Time Series Based on Fuzzy Theory and Genetic Algorithm, Physics Procedia, Volume 33, 2012, Pages 1241-1247, ISSN 1875-3892,

[36] Yahoo (2014) Yahoo Inc stock data YHOO [Online] Available at < https://uk.finance.yahoo.com/q/hp?s=YHOO> [Accessed: 20/06/2014]

[37] Ye F., Mabu S., Wang L., Hirasawa K. (2009) Genetic Network Programming with General Individual Reconstruction ICROS-SICE International Joint Conference 2009 August 18-21, 2009, Fukuoka International Congress Center, Japan

[38] Zhang X., Onieva E., Perallos A., Osaba E., Lee V. C. S. (2014) Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction, Transportation Research Part C: Emerging Technologies, Volume 43, Part 1, June 2014, Pages 127-142, ISSN 0968-090X,