DE GRUYTER OPEN

Scientific Annals of the "Alexandru Ioan Cuza" University of Iaşi Economic Sciences 62 (1), 2015, 55-62 DOI 10.1515/aicue-2015-0004

GENERATING INVESTMENT STRATEGIES USING MULTIOBJECTIVE GENETIC PROGRAMMING AND INTERNET TERM POPULARITY DATA

Martin JAKUBÉCI*

Abstract

Searching for stock picking strategies can be modelled as a multiobjective optimization problem. The objectives are mostly the profit and risk. Because of the conflicting nature of these objectives, we have to find Pareto optimal solutions. Multiobjective genetic programming (MOGP) can be used to find tree based solutions, using evolutionary operators. The advantage is that this algorithm can combine any number of inputs and generate complex models. Recent research shows, that the popularity of different terms on the internet can be used to enhance the models. This paper deals with a SPEA2 MOGP implementation, which uses Google trends and Wikipedia popularity to find stock investment strategies.

Keywords: genetic programming, Google trends, stock

JEL classification: G11

1. INTRODUCTION

Financial markets are complex systems, which consist of many interacting entities. That's why they are hard to predict. Investors are trying to create portfolios of assets to achieve high profit and minimize the risk. Low prices of hardware and data availability caused high interest in computer modelling in the area of investing. Popular group of algorithms, that are used for modelling are the evolutionary algorithms. One of them is genetic programming, which uses operators inspired by the evolution theory to generate tree programs. These programs can represent stock picking strategies.

2. RELATED RESEARCH

There is a lot of research in the area of stock picking using multiobjective genetic programming, for example (Allen and Karjalainen, 1999, p. 21; Lohpetch and Corne, 2011; Skolpadungket *et al.*, 2007; Bradshaw *et al.*, 2009; Hassan, 2010 and Chen *et al.*, 2014). But there is no consensus on whether the algorithm is able to outperform the market and secure high revenues (Chen and Navet, 2007, p. 1; Potvin *et al.*, 2004, p. 14).

* Department of Information Systems, Faculty of Management, Comenius University in Bratislava, Slovak Republic; e-mail: *martin.jakubeci@gmail.com.*

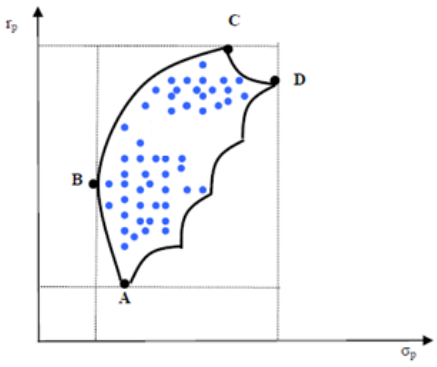
Internet term popularity in the financial area is researched only recently. Most of the research always concentrated on historical prices and different technical and fundamental indicators. Big interest was caused by the article dealing with market index changes caused by Google trends changes of different terms (Preis *et al.*, 2013). Similar research was done with page views on Wikipedia (Moat *et al.*, 2013), terms in Facebook statuses (Karabulut, 2013) and Twitter posts (Ruiz *et al.*, 2012). None of this research used evolutionary algorithms.

3. FINANCIAL MARKETS

There are multiple models and strategies. Efficient market hypothesis believes, that financial markets are effective and every information is immediately absorbed. The price of the asset reflects all information and is equal to the real value (Beechey *et al.*, 2000, p. 2). Changes in price are random fluctuations around this value and can be described as a random walk (Alexander, 2008, p. 134).

When investing, there is always some risk. Modern investment strategies are based on diversification, investing in multiple assets (Bohdalová and Greguš, 2011, p. 2). This was formalized by Harry Markowitz as multiobjective problem, where the objectives are revenue maximization and risk minimization. Investment strategy is a rule, which specifies investor's position on every asset in time *t*. The position is chosen based on the available information, without the knowledge of the future (Bohdalová and Greguš, 2012, p. 21).

Risk and revenue values are assigned to every portfolio, this can be seen on Figure 1. Blue dots are the available portfolios and portfolios on the line between B and C are the optimal strategies, or Pareto optimal solutions. This means that value of none of the objectives can be increased without sacrificing value of a different objective. That means, that they are Pareto dominant over other solutions and build the Pareto front (Hassan, 2010, p. 10).



Source: Toman (2008, p. 18)

Figure no. 1 – Portfolio revenue and risk

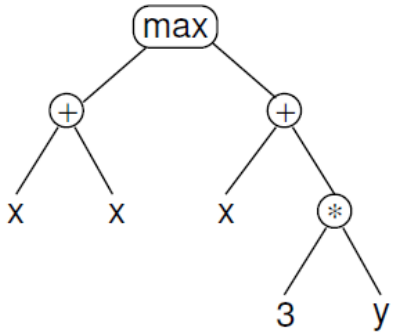
Investors opposing the efficient market hypothesis believe, that the market can be outperformed and high returns can be achieved. Many strategies are based on the fundamental analysis. Its main idea is that real value of a stock and its price on the market can differ and it should be invested into undervalued companies. Finding the real value is not an easy task, it requires analysis of financial and other data (Thomsett, 2006, p. 2).

Other group of strategies is based on the technical analysis. The future prices are predicted from the historical prices. It is based on three principles. The first one is that the price reflects everything, the second one is that prices move in trends and the last one is that history repeats (Chovancová, 2006, p. 315).

4. GENETIC PROGRAMMING

Genetic programming is an evolutionary optimization algorithm, which is searching for problem solutions. Solution is a program represented by a tree structure. First generation of solutions is created randomly. Every next generation is created by stochastic transformation of the previous generation. Transformation is done by applying operators, which are inspired by the evolution theory. These operators are mostly selection, mutation and crossover. Every next generation is expected to be better, the quality of the solutions is evaluated by the fitness function (Poli *et al.*, 2008, p. 2).

The most frequently used representation of a solution is a syntactic tree. The solution is in fact a program, which can be split into commands, organized as a tree. Example of such tree can be seen on Figure 2. Programs can be represented also in the prefix form, which is known from the functional programming. max(x+x,x+3*y) is written as (max (+ x x) (+ x (* 3 y))). The relation between commands and subcommands is more obvious in the prefix form.



Source: Poli et al. (2008, p. 10)

Figure no.2 – Tree based representation of a max(x+x,x+3*y) program

The tree based solutions are formed from 2 different sets of vertices. The first group are terminal symbols, for example inputs, contents or any method calls, which do not accept any parameters. Those are leafs of the tree structure. The second set are nonterminals, or methods, that accept parameters. For example arithmetic operators, logical operators, conditions etc. They are expected to be type and run safe, so that the solutions can be executed to transform inputs to outputs. The first vertex in the tree is called root and the depth of every vertex is defined as the distance from the root.

First step during the run of the genetic programming is the initialization of the population (the set of the solutions). This done by creating random solutions. Two common methods are the full method, where every leaf has the same depth, equal to the maximum depth specified for the solutions. The other method is the rising method, where terminals and nonterminals are applied randomly and the depths are different, but never more than the

maximum depth. Maximum depth is important to avoid very large trees, which are too time consuming.

The other generations are created by applying genetic operators. Part of the new generation is created by copying best solutions from the previous generation. The other operator is the mutation, where a random subtree or leaf is replaced by a random subtree or leaf. Crossover is similar, but two solutions exchange their random subtrees or leafs (Poli *et al.*, 2008, pp. 9-27).

As already mentioned, the quality of the solution is evaluated by the fitness function. Solution is filled with inputs, executed and the output is evaluated. When dealing with multiobjective optimization, there are multiple fitness functions required, one for every objective. There are many algorithms to handle multiple objectives in evolutionary algorithms. SPEA2 was chosen, because it overcomes some issues in other algorithms. It's based on elitism, Pareto dominant solutions are kept in a separate archive with fixed size (Hassan, 2010, p. 20). The algorithm works this way (Zitzler *et al.*, 2001, p. 5):

Input: *N (population size)*
M(archive size)
T (maximum number of generations)
Output: ***A*** *(non-dominated set)*

Step 1: ***Initialization****: Generate an initial population and create the empty archive (external set);. Set t* = 0.
Step 2: ***Fitness assignment****: Calculate fitness values of individuals in population and archive.*
Step 3: ***Environmental selection****: Copy all non-dominated individuals in population and archive to the new archive. If size of the new archive exceeds M then reduce new archive by means of the truncation operator, otherwise if size of new archive is less than N then fill new archive with dominated individuals in population and archive.*
Step 4: ***Termination****: If t >=T or another stopping criterion is satisfied then set* ***A*** *to the set of decision vectors represented by the non-dominated individuals in the archive. Stop.*
Step 5: ***Mating selection****: Perform binary tournament selection with replacement on the new archive in order to fill the mating pool.*
Step 6: ***Variation****: Apply recombination and mutation operators to the mating pool and set new population to the resulting population. Increment generation counter (t =t + 1) and go to Step 2.*

5. GOAL

Goal of this research is to implement a multi-objective genetic programming algorithm, which uses internet term popularity data to find investment strategies. These strategies are then compared with the buy and hold strategy.

6. METHODS

The algorithm is searching for portfolio creating strategies, which return a stock ranking based on the inputs. The strategy is evaluated every day for the specified training period. When the ranking is in the bottom third and the stock is in portfolio, it is sold. If it is in the upper third, it is bought. The average yearly revenue percentage and standard deviation of revenues is evaluated in the training data. These two values are calculated by the two fitness methods. Training was done on data in the years 2010-2013. Starting capital is 100000 USD and the stocks in the portfolio are chosen every day in training period. Maximum tree depth was set to 10. Google popularity was used for the name of the company and Wikipedia popularity of the article about the company that is represented by the evaluated stock.

Data for the 30 Dow Jones Industrial Index companies was used. During evaluation, this data is available for the strategy:

- the list of opening, closing, lowest and highest daily price in the last 50 days,
- opening, closing, lowest and highest daily price on the last day,
- Google popularity of company name in the last 50 days,
- Google popularity of company name on the last day,
- Wikipedia company page popularity in the last 50 days,
- Wikipedia company page popularity on the last day.

Historical prices were downloaded from Yahoo Finance, Google term popularity from Google Trends and Wikipedia article popularity from Wikipedia article traffic statistics. These functions (non-terminals) were used:

- arithmetic operations: addition, subtraction, multiplication, division, negation and exponentiation,
- logical operations: conjunction, disjunction, negation,
- equality: higher, lower, equal, or any combination
- trigonometric operations: sine, cosine,
- condition,
- list operations: lag, moving average.

Implementation was done in the C# language using expression trees. They allow working with an algorithm as a data structure, so modifications of the solutions and application of the evolutionary operators is possible. The Metalinq (2014) library was used, to simplify these modifications.

We used rate of return to compare genetic programming, which is calculated as the difference between the portfolio value at end and portfolio value at the beginning, divided by the portfolio value at the beginning.

7. RESULTS

Distribution of the strategies can be seen on Figure 3, revenue is on the y axis and standard deviation on the x axis. The Pareto front can be seen in the upper left area.

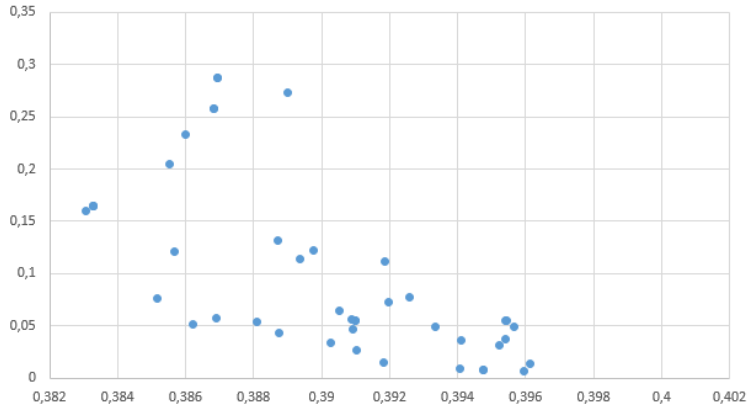


Figure no.3 – Portfolio revenue and deviation

Sample investment strategy looks like this:

```
((lastOpen + -Invoke((l, c) =>l.Skip(c).FirstOrDefault(), open, 5)) * (((IIF((lastWiki>= lastOpen), (Invoke((l, c) =>l.Skip(c).FirstOrDefault(), close, 1) / lastGoogle), Invoke((l, c) =>l.Skip(c).FirstOrDefault(), wiki, 42)) / Invoke((l, c) =>l.Skip(c).FirstOrDefault(), high, 22)) + (Invoke((l, c) =>l.Skip(c).FirstOrDefault(), Invoke((list, period) =>list.Skip((period - 1)).Select((item, index) => (list.Skip(index).Take(period).Sum() / Convert(period))).ToList(), Invoke((list, period) =>list.Skip((period - 1)).Select((item, index) => (list.Skip(index).Take(period).Sum() / Convert(period))).ToList(), google, 5), 5), 5) ^ Invoke((l, c) =>l.Skip(c).FirstOrDefault(), open, 46))) ^ (Invoke((l, c) =>l.Skip(c).FirstOrDefault(), high, 22) / IIF(((lastClose<lastVolume) And (lastOpen>= 0,16)), Invoke((l, c) =>l.Skip(c).FirstOrDefault(), Invoke((list, period) =>list.Skip((period - 1)).Select((item, index) => (list.Skip(index).Take(period).Sum() / Convert(period))).ToList(), Invoke((list, period) =>list.Skip((period - 1)).Select((item, index) => (list.Skip(index).Take(period).Sum() / Convert(period))).ToList(), google, 40), 13), 39), (Invoke((l, c) =>l.Skip(c).FirstOrDefault(), open, 1) * Invoke(l => Cos(l), lastClose))))))
```

The method *(list, period) =>list.Skip(period - 1).Select((item, index) =>list.Skip(index).Take(period).Sum() / period).ToList()* is a moving average on the *list* with the time window of size *period*. The method (l, c) =>l. Skip(c).First Or Default() is the lag of the list*l*of size *c*.

It can be seen that the strategy is quite complicated and it represents a nonlinear model with many variables and functions. Google and Wikipedia data is present on multiple places.

Revenues from this strategy are compared to the market index on Figure 4 (training set 2010-2013) and on Figure 5 (evaluation set 2014).

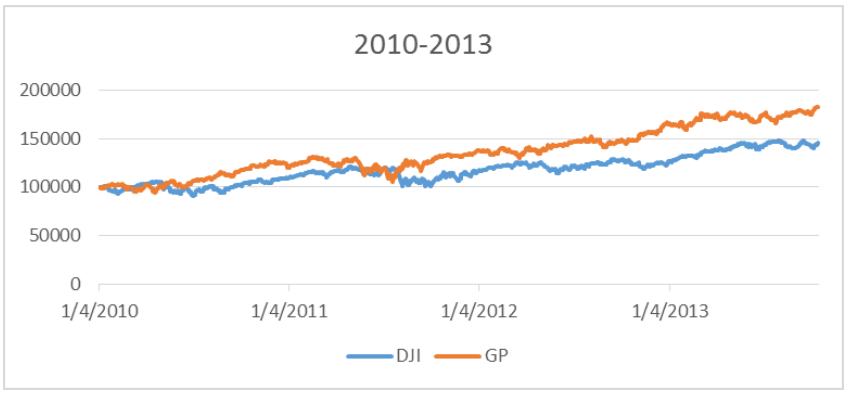


Figure no.4 – Value of the portfolios on the training data

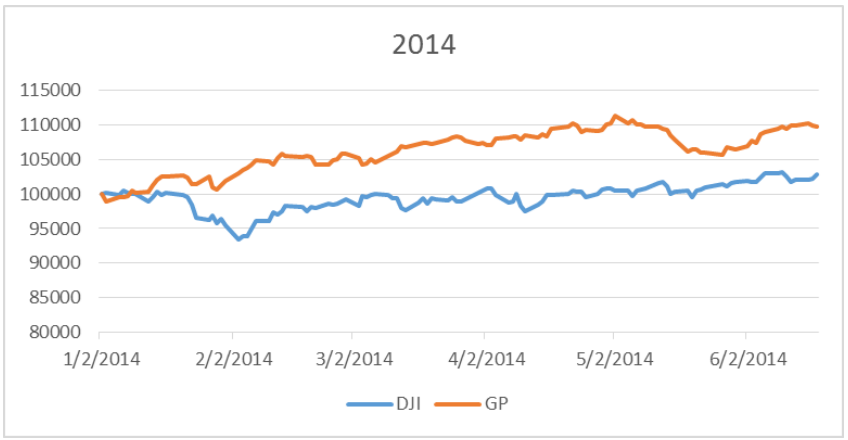


Figure no. 5 – Value of the portfolios outside of the training data

It can be seen, that the genetic programming trained strategy outperforms the market. Genetic programming is better in almost whole period, with some small temporary exceptions.

8. CONCLUSION

This paper dealt with generating investment strategies using genetic programming. Data about historical prices and internet term popularity data from Google and Wikipedia was used as input. The implementation was compared with the buy and hold strategy on the DJI index.

It was shown that the implementation is able to outperform the market index, even outside of the training data. To compare the performance, rate of return was used. This has proven the usability of this implementation and further research should be done.

In future, this implementation should be compared with other investment strategies and also on more periods.

References

Alexander, C., 2008. *Quantitative methods in finance*. Chichester: John Wiley and Sons, Ltd.

Allen, F., and Karjalainen, R., 1999. Using genetic algorithms to find technical trading rules. *Journal of Financial Economics, 51*, 245–271.

Beechey, M., Gruen, D., and Vickery, J., 2000. The efficient market hypothesis: a survey: Reserve Bank of Australia,.

Bohdalová, M., and Greguš, M., 2011. The identification of key market risk factors for a portfolio of EU bonds. *Global business and economics anthology, 2*(2), 470-477.

Bohdalová, M., and Greguš, M., 2012. Portfolio optimization and Sharpe ratio based on copula approach. *Research Journal of Economics, Business and ICT, 6*, 6-10.

Bradshaw, N. A., Walshaw, C., Ierotheou, C., and Parrott, A. K., 2009. A Multi-Objective Evolutionary Algorithm for Portfolio Optimisation. *Proceedings of the Adaptive and Emergent Behaviour and Complex Systems Convention*, 27–32.

Chen, S. H., and Navet, N., 2007. Failure of Genetic-Programming Induced Trading Strategies: Distinguishing between Efficient Markets and Inefficient Algorithms. *Computational Intelligence in Economics and Finance, 2*, 169-182.

Chen, S. S., Huang, C. F., and Hong, T. P., 2014. An Improved Multi-Objective Genetic Model for Stock Selection with Domain Knowledge. *Technologies and Applications of Artificial Intelligence, Lecture Notes in Computer Science, 8916*, 66-73.

Chovancová, B., 2006. *Finančnýtrh – nástroje, transakcieainštitúcie*. Bratislava: Iura Edition.

Hassan, G. N. A., 2010. *Multiobjective genetic programming for financial portfolio management in dynamic environments.* (Doctoral thesis), University College London.

Karabulut, Y., 2013. *Can Facebook Predict Stock Market Activity?* , Goethe University Frankfurt.

Lohpetch, D., and Corne, D., 2011. Multiobjective algorithms for financial trading: Multiobjective out-trades single-objective. *IEEE Congress on Evolutionary Computation*, 192–199.

Metalinq, 2014. MetaLinq - LINQ to Expressions. from http://metalinq.codeplex.com/

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., and Preis, T., 2013. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports, 3*.

Poli, R., Langdon, W. B., and McPhee, N. F., 2008. A Field Guide to Genetic Programming. Retrieved 25 March, 2014, from http://www.gp-field-guide.org.uk

Potvin, J. Y., Soriano, P., and Vallée, M., 2004. Generating trading rules on the stock markets with genetic programming. *Computers & Operations Research, 31*(7), 1033–1047.

Preis, T., Moat, H. S., and Stanley, H. E., 2013. Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports, 3*.

Ruiz, J. E., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A., 2012. *Correlating Financial Time Series with Micro-Blogging Activity*. Paper presented at the WSDM'12, Seattle, WA.

Skolpadungket, P., Keshav, D., and Harnpornchai, N., 2007. Portfolio Optimization using Multi-objective Genetic Algorithms. *IEEE Congress on Evolutionary Computation*, 516 - 523.

Thomsett, M. C., 2006. *Getting started in fundamental analysis*. Hoboken: John Wiley & Sons, Inc.

Toman, R., 2008. *Analýza faktorového portfolio najviac likvidných cenných papierov na BCCP v závislosti na HDP, inflácii a PX*. (Master's thesis), Masaryk University, Brno.

Zitzler, E., Laumanns, M., and Thiele, L., 2001. SPEA2: Improving the Strength Pareto Evolutionary Algorithm. *Evolutionary Methods for Design, Optimization, and Control*, 95–100.