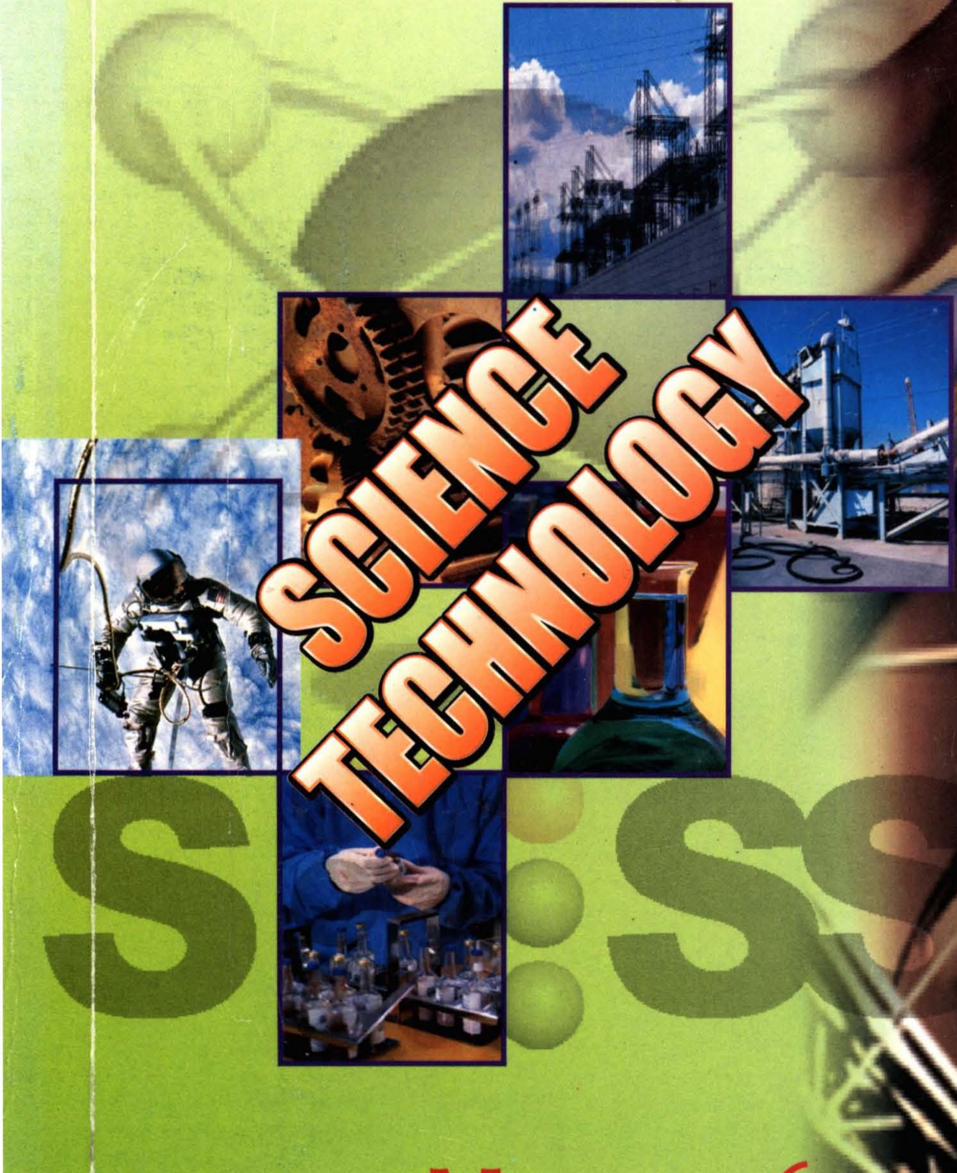
UNIVERSITI
TEKNOLOGI
MARA

ISO 9001:2000 No Sijil :0500132

NATIONAL SEMINAR ON SCIENCE TECHNOLOGY & SOCIAL SCIENCES 2006

Globalising Knowledge and Information



30-31 May 2006

Swiss Garden Resort & Spa
Kuantan, Pahang

Data Mining using Genetic Algorithm in Finance Data

A. Noor Latiffah
A. B. Nordin

ABSTRACT

Computing systems has enabled us to collect tremendous amount of data and information. A large pool of data requires not only an efficient and effective retrieval system but also a better way to discover hidden knowledge. Data mining can discover patterns or rules from a vast volume of data. This patterns or rules may help to develop better decision-making process. Data mining is primarily used in finance and business environment to extract knowledge from financial, retail, communication and marketing data. This project, will extract some useful financial knowledge from the Syariah Index data of Kuala Lumpur Syariah Index (KLSI). The methods that will be applied are conventional statistical methods: Markowitz Optimization as well as evolutionary programming (EP) utilizing genetic algorithms. The result of this project are expected to be a comparison of the used methods that will give an indication how well evolutionary programming can perform relative to conventional method and how good the results of the data mining process.

Keywords: *Data Mining, Genetic Algorithms, Markowitz Optimization*

Introduction

Human analysts with no special tools can no longer make sense of enormous volumes of data that require processing in order to make informed business decisions. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can be either utilized in an automated decision support system or assessed by a human analyst. Can one predict the most profitable securities to buy/sell during the next trading session? This kind of question can probably be answered if information hidden among megabytes of data in our database can be found explicitly and utilized.

Genetic Algorithms

Genetic Algorithms (GA) is a heuristic function for optimization, where the extreme of the function (i.e., minimal or maximal) cannot be established analytically. A population of potential solutions is refined iteratively by employing a strategy inspired by Darwinist evolution or natural selection. Genetic Algorithms promote survival of the fittest. This type of heuristic has been applied in many different fields, including construction of neural networks and finance. We represented the parameters of a trading rule with a one-dimension vector that is called chromosome, each element is called a gene, and all of the chromosomes are called population. Here, each gene stands for a parameter value; each chromosome is the set of parameters of one trading rule. Generally, genetic operations include: crossover, mutation and selection. (Lin *et al.* 2004)

Markowitz Portfolio Optimization

The portfolio model introduced by Markowitz (1959) assumes an investor has two considerations when constructing an investment portfolio: expected return and variance in return (i.e., risk). Variance measures the variability in realized return around the expected return, giving equal weight to realizations below the expected and above the expected return. The Markowitz model requires two major kinds of information: (1) the estimated expected return for each candidate investment and (2) the covariance matrix of returns. The covariance matrix characterizes not only the individual variability of the return on each investment, but also how each investment's return tends to move with other investments.

Why Markowitz Portfolio Optimization and GA?

Even though GA is not new in artificial intelligence, the application of GA in data mining is somehow lower than other methods (Rexer 2003). Lin *et al.* (2004) has approved that GA can be used in stock market optimization. The change in data environment has challenged the GA to take part in data mining process.

Markowitz Portfolio Theory is a well-known technique. In March 1952, issue of Journal of Finance, has published an article from Harry M. Markowitz called "Portfolio Selection". In it, he demonstrated how to reduce the standard deviation of returns on asset portfolios by selecting assets, which do not move in exactly the same ways. At the same time, he laid down some basic principles for establishing an advantageous relationship between risk and return, and his work is still in use forty years later. For the purpose of comparisons, Markowitz is a strong benchmark to look for.

Methodology

Ishak, Nordin, and Salwana (2002) have done a research using Markowitz Portfolio Optimization for two Syariah equities with equal weight from the consumer sector of main board of Bursa Malaysia. In this project, a simple Markowitz Portfolio Optimization is used in order to select the best portfolio. To carry out the simple Markowitz Portfolio Optimization, the following statistics has been calculated:

i) the mean return for equity i,

$$R_i = \frac{1}{N}\sum_{t=1}^{N} R_{it},\ i = 1,2,3,\ldots,N$$

where is the return of the t th week

N is the total number of week in this study.

ii) the standard deviation of return for the equity i is

$$\sigma_i = \left[\frac{1}{N}\sum_{t=1}^{N}(R_{it} - R_i)^2\right]^{\frac{1}{2}},\ i = 1,2,3,\ldots,n$$

where n is the number of equities.

iii) the expected return of the pth portfolio is

$$E(R_p) = \sum_{i=1}^{n} W_i R_i,\ p = 1,2,3,\ldots,k$$

where is the proportion of the ith equity in the portfolio p

is the mean return of the ith equity
k is the number of portfolio.

iv) the standard deviation of the pth portfolio is

$$\sigma_p = \left[\sum_{i=1}^{N} W_i^2\sigma_i^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} W_i W_j\ COV(i,j)\right]^{\frac{1}{2}}$$

where $COV(i,j)$ is the covariance of the ith and jth equities.

v) utility portfolio p is

$$E(U_p) = E(R_p) - RP$$

where

RT is a risk tolerance (0-100%) which reflects the investor's willingness to bear more risk for more return.

vi) the optimal (best) portfolio for an investor would be the one from the opportunity set that maximizes utility.

To apply the GA, the following operation has been included in the process.

i- Crossover operator.

Supposed are two chromosomes, select a random integer number expect are offspring of

$$S_3 = \{s_i \mid if\ i \le r,\ s_i \in S_1,\ else\ s_i \in S_2\},$$
$$S_4 = \{s_i \mid if\ i \le r,\ s_2 \in S_1,\ else\ s_i \in S_1\}$$

crossover ( ) selects two parents that took part in the cross over where we are implementing a single point crossover.

ii- Mutation operator.

Supposed a chromosome select a random integer number expect, is a mutation of In this study, we are using random uniform mutation. A variable selected for mutation is replaces by a random value between the expected return and its variance.

iii- Selection operator.

Supposed there are *m* individuals, we select [*m*/2] individuals but erase the others, the ones we selected are "more fitness" that means their profits are greater. We are using the standard proportional selection for maximizing problems incorporating elitist model to make sure that the best member survives. The following algorithm outlines the selection process.

```
selection( )
{
find the total fitness of the population
          calculate relative fitness
          calculate cumulative fitness
select survivors using cumulative fitness
}
```

Figure 1 explains the process involved.

$E(R_p)$	σ_p^2	RT

For this study, we have identified the chromosomes or the candidate solutions to be as follows:

The objective function is

$$f(u_p) = E(R_p) - \left(\frac{\sigma_p^2}{RT}\right),$$

where expected return of the *p*th portfolio, is variance of return of *p*th portfolio and is a risk tolerance.

For this particular problem, we have used the following parameters: population size, POPSIZE = 505, probability of crossover, PXOVER = 0.8, and probability of mutation, PMUTATION = 0.15.

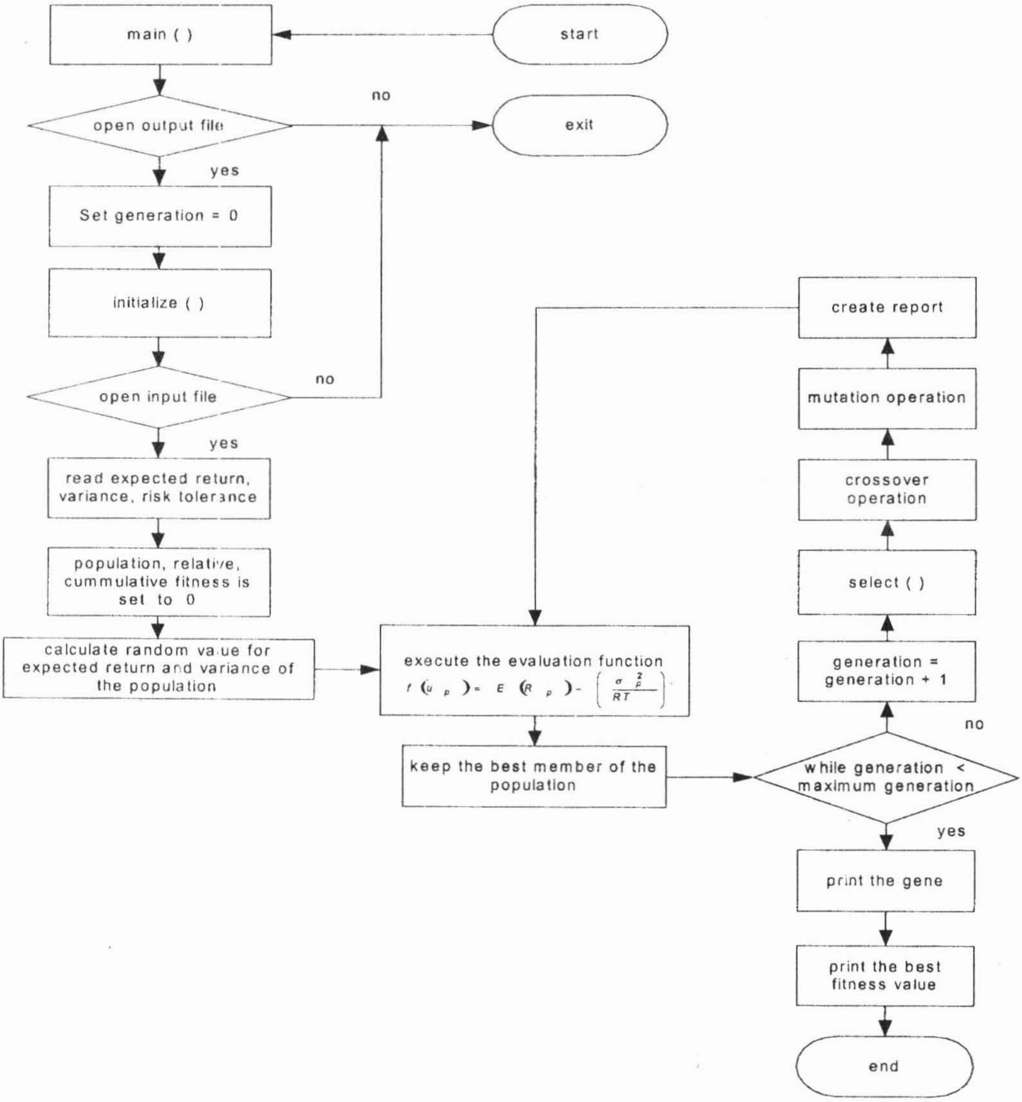

main ( )
start
open output file
no
exit
yes
Set generation = 0
initialize ( )
open input file
no
yes
read expected return, variance, risk tolerance
population, relative, cummulative fitness is set to 0
calculate random value for expected return and variance of the population
execute the evaluation function
$f(u_p) = E(R_p) - \left(\frac{\sigma_p^2}{RT}\right)$
keep the best member of the population
while generation < maximum generation
no
yes
generation = generation + 1
select ( )
crossover operation
mutation operation
create report
print the gene
print the best fitness value
end

Fig. 1: Genetic Algorithm Flowchart

Results and Discussion

The GA program was coded using C programming language. Using the program, the utility has been calculated through the objective function , $f(u_p) = E(R_p) - \left(\frac{\sigma_p^2}{RT}\right)$. The summary of the result is shown in Table 1 together with utilities calculated using Markowitz Portfolio Optimization.

Table 1: Utilities of all possible portfolios with various risk tolerance.

Portfolio	10% Risk		30% Risk		50% Risk		70% Risk		100% Risk	
	Markowitz	GA	Markowitz	GA	Markowitz	GA	Markowitz	GA	Markowitz	GA
P1	0.0964	0.257	0.1063	0.2572	0.1082	0.1246	0.1091	0.2374	0.1097	0.2067
P2	0.0862	0.0015	0.0966	0.0000	0.0987	0.1597	0.0996	0.1210	0.1002	0.0342
P3	0.1665	0.2159	0.1796	0.0993	0.1822	0.0416	0.1833	0.1460	0.1841	0.3702
P4	0.2066	0.2677	0.2163	0.1014	0.2182	0.1992	0.219	0.1743	0.2196	0.3647
P5	0.1236	0.0149	0.1288	0.0432	0.1298	0.1634	0.1303	0.023	0.1306	0.2059
P6	0.2035	0.3266	0.2116	0.0115	0.2133	0.0598	0.214	0.1657	0.2145	0.1707
P7	0.2454	0.2858	0.2505	0.2551	0.2496	0.0091	0.2499	0.2184	0.2502	0.1515
P8	0.2217	0.2761	0.2348	0.4192	0.2374	0.2897	0.2385	0.3934	0.2393	0.1392
P9	0.2362	0.1737	0.2396	0.2151	0.2403	0.2612	0.2405	0.2836	0.2407	0.2064
P10	0.3148	0.2808	0.322	0.2157	0.3234	0.0382	0.3241	0.2604	0.3245	0.3637

The graph for each risk tolerance is plotted for Markowitz and GA. For 10% , 30% and 70% risk tolerance, Markowitz and GA follow the same route for utility with up and down trend. For 50% risk tolerance, GA utility is declining obviously over Markowitz for P6 and P7. For 100% risk tolerance, GA move opposite to Markowitz especially for P7 and P8.

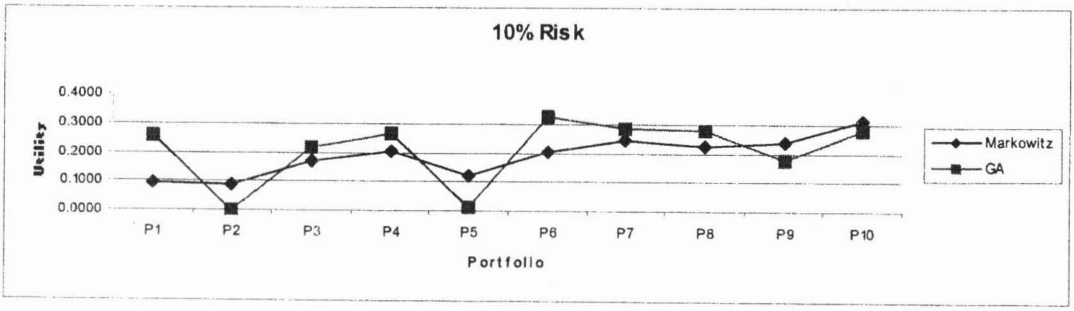


Fig. 2: 10% Risk tolerance

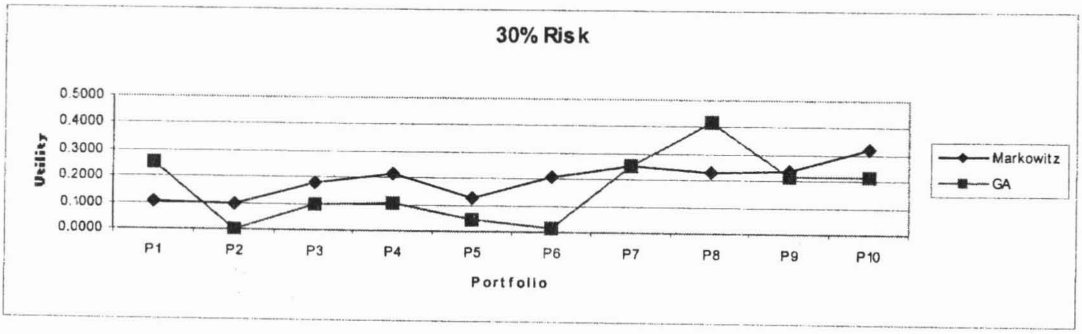


Fig. 3: 30% Risk tolerance

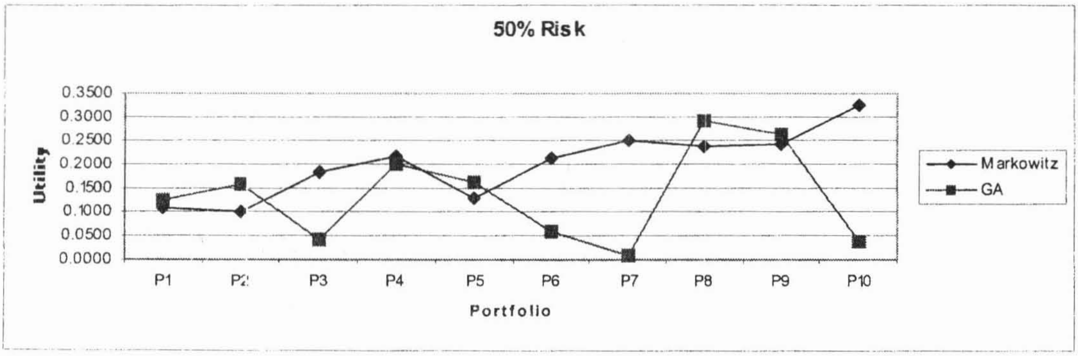


Fig. 4: 50% Risk tolerance

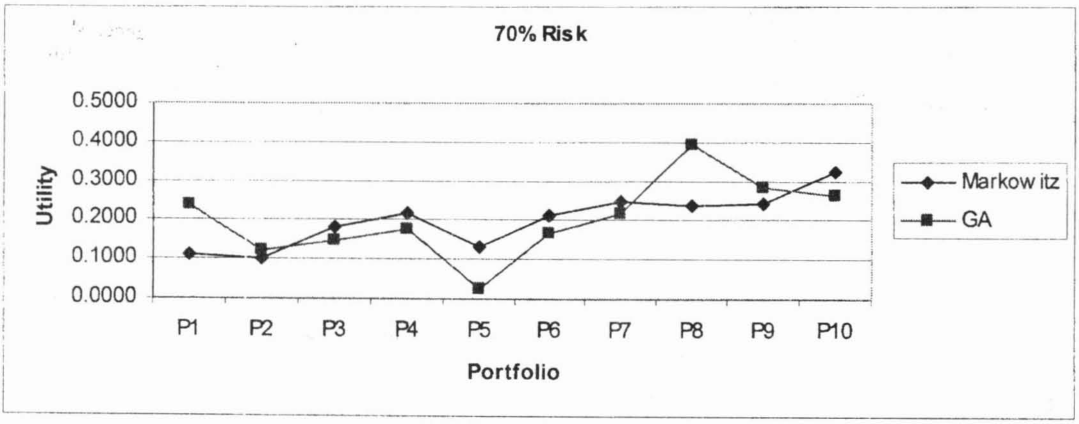


Fig. 5: 70% Risk tolerance

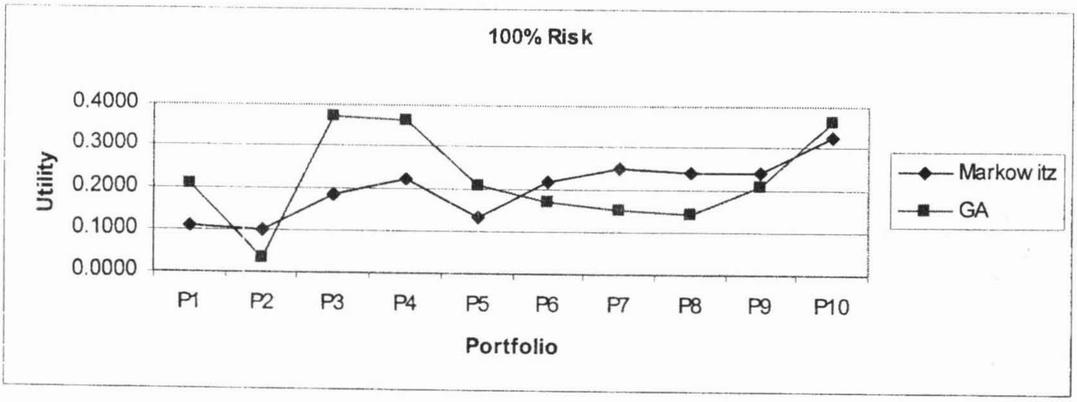


Fig. 6: 100% Risk tolerance

By using Markowitz Portfolio Optimization, it is clear that, the optimum portfolio is P10, which consist of UMW and Nestle. The next step is to identify which one is the best among the ten portfolios in terms of the utilities using GA. Graphs between utilities and risk tolerances are drawn for groups of portfolios. The graphs are as follows and shown in Figures 7 -10.

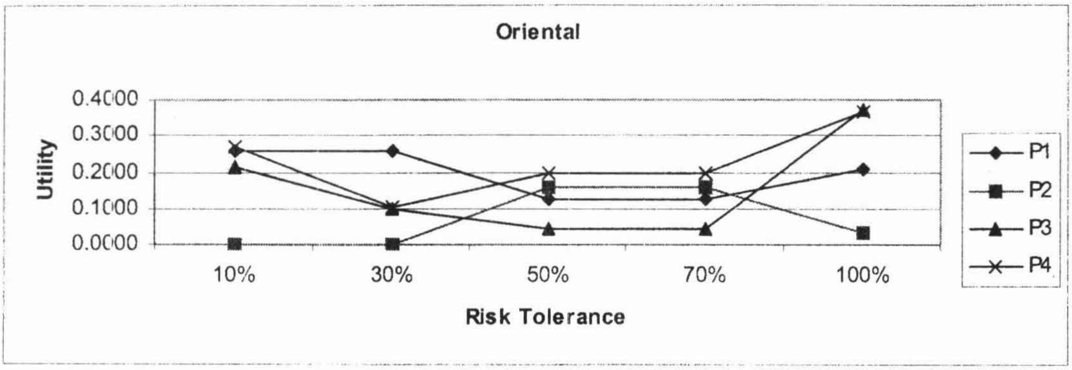


Fig. 7: Utility of selected portfolio: Oriental always invested at 50% with the other four securities

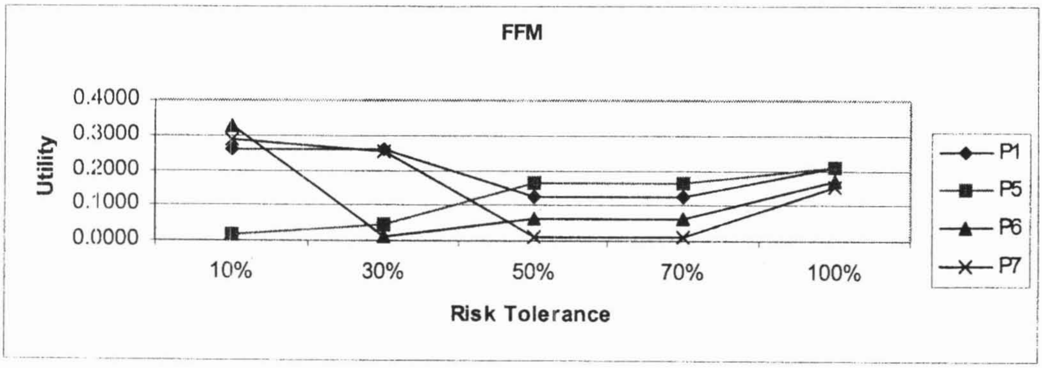


Fig. 8: Utility of selected portfolio: FFM always invested at 50% with the other four securities

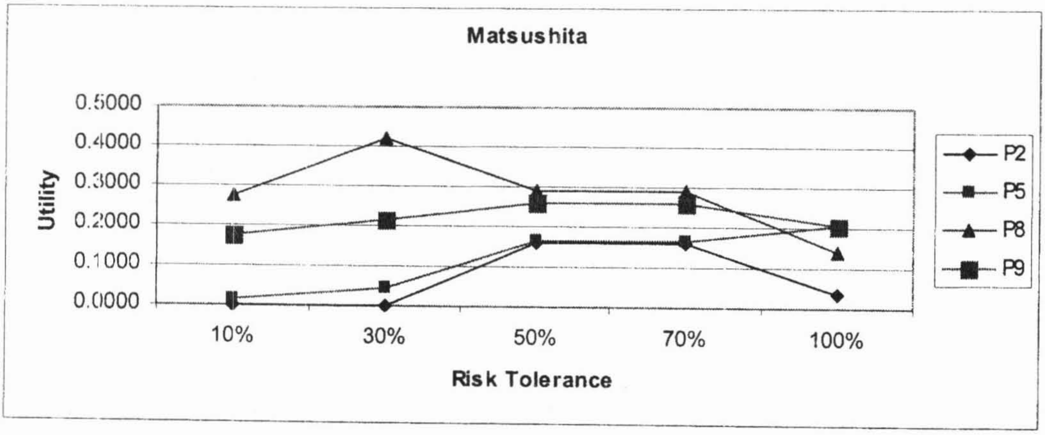


Fig. 9: Utility of selected portfolio: Matsushita always invested at 50% with the other four securities

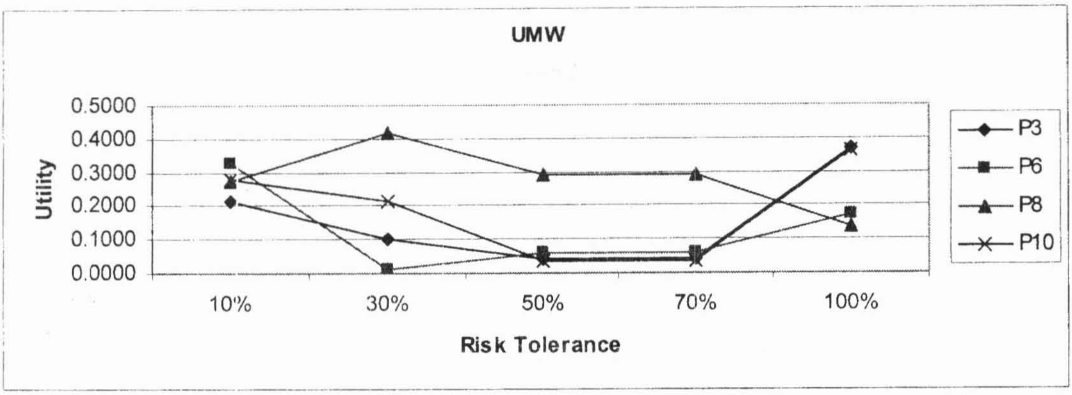


Fig. 10: Utility of selected portfolio: UMW always invested at 50% with the other four securities

From the graph, the best utilities among the best portfolios is identified and shown in Table 2. The graph is plotted to get a clearer view in Figure 11.

Table 2: The best utilities

Portfolio	10% Risk	30% Risk	50% Risk	70% Risk	100% Risk
P1	0.2570	0.2572	0.1246	0.1246	0.2067
P8	0.2761	0.4192	0.2897	0.2897	0.1392

It is clear that P8 which consist of Matsushita with UMW is the best since it has the highest utilities for all tolerance.

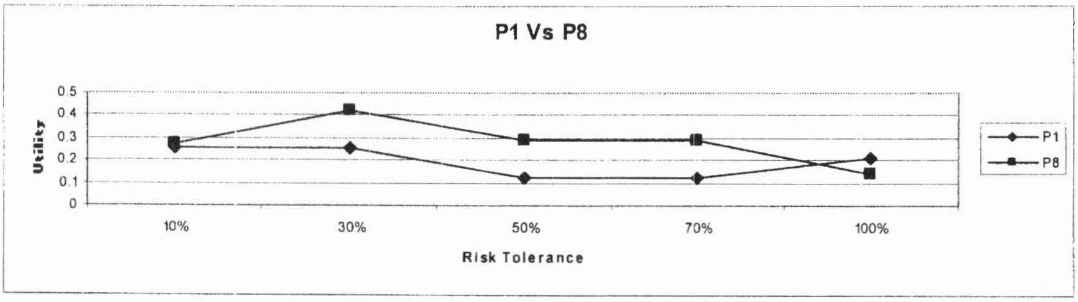


Fig. 11: The best utilities

GA is a heuristic function for optimization, where the extreme of the function (i.e., minimal or maximal) cannot be established analytically. A population of potential solutions is refines iteratively by employing a strategy of natural selection. The iterative process continues until one of the possible termination criteria is met: if a known optimal or acceptable solution level is attained; or if a given number of generations without fitness improvement occur or in our case, if a maximum number of generations have been performed. GA promotes the 'survival of the fittest' (Lin et al. 2004). The difference in the result may be because of the random value generated by the program and our selection of objective function. A study by Foster and Meysenberg (1999) shows that random value does affect GA performance.

GA particularly applicable to problems, which are large, non-linear and possibly discrete in nature, features that traditionally; add to the degree of complexity of solution. Population size selection is probably the most important parameter, reflecting the size and complexity of the problem. However, the trade-off between extra computational effort with respect to increase population size is a problem specific decision to be ascertained by the modeler, as doubling the population size will approximately double the solution time for the same number of generations.

Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. This raises the issues of scalability and efficiency of data mining methods when processing considerably large data. GA is suitable for mining financial data such as stock market.

Conclusions

The optimum portfolio generated by GA differs from optimum portfolio generated by Markowitz Optimization. Markowitz optimization produces UMW and Nestle as the best portfolio of equal percentage allocation for any risk tolerance. GA on the other hand, generated a combination of UMW and Matsushita as the optimum portfolio. Markowitz Portfolio Optimization is an established principle in the investment sector as his work is still in use for more than four decades. However, we are dealing with new technologies and massive data, it is best to find a new way to implement the Markowitz principle. GA using Markowitz optimization in this study has shown that it is better than Markowitz alone. We have to consider the running time and also data volume to be processed in order to make an accurate decision. We are mining the data to make better investment with higher profit and lower risk.

References

Ishak, A. G., Nordin, A. B. and Salwana, H. (2002). *Markowitz Portfolio Optimization for Two Syariah Equities with Equal Weight from the Consumer Sector of Main Board.* Kolokium Penyelidikan.

Markowitz, H. M. (1952). The Portfolio Selection. *The Journal of Finance.* [online]. Available: http://cowles.econ.yale.edu/P/cp/p00b/p0060.pdf

Shoaf, J.S. and Foster, J.A. (1996) "*A Genetic Algorithm Solution to the Efficient Set Problem: A Technique for Portfolio Selection Based on the Markowitz Model.*" Proceedings of the 1996 Annual Meeting, Decision Sciences Institute. Vol. II. p 571-573.

Shoaf, J.S. and Foster, J.A. (1998). "*The Efficient Set GA for Stock Portfolios*". Proceeding International Conference on Evolutionary Computing (CEC). IEEE Press.

Rexer, K. (2003). Data mining techniques (Nov 2003). [Online]. Available: http://www.kdnuggets.com/polls/2003/data_mining_techniques.htm

Lin *et al.* (2004). The applications of Genetic Algorithms in Stock Market Data Mining Optimisation. [Online]. Available: http://www-staff.it.uts.edu.au/~lbcao/publication/DM2004.pdf

A. NOOR LATIFFAH, Universiti Teknologi MARA Shah Alam.

A. B. NORDIN, Universiti Teknologi MARA Shah Alam.