

# Application of Linkage Learning Genetic Algorithm on Stock Selection – Evidence from Indian Stock Market

R. Lakshmi<sup>1</sup>, and Dr. S. Amilan<sup>2</sup>

<sup>1</sup>Assistant professor, Department of Computer Science, Pondicherry University, India

E-mail id: rlakshmiseYlva@yahoo.co.in

<sup>2</sup>Associate Professor, Department of CommIerce, Pondicherry University, India

E-mail id: amilan.kcm@pondiuni.edu.in

**Abstract:** *The financial world is interested in selecting the stocks which are superior in performance in the given group of stocks. For this purpose, this paper uses a new Linkage Learning method which draws its concept from machine learning approach and incorporated it with the existing Genetic Algorithms. The proposed method finds a stock with better performance among the given number of stocks using historical/fundamental financial indicators and price information of stocks that can be graded for its performance. Experimental results with real data from the Indian Stock Market reveal that the proposed method used in the present study for stock selection leads to better results than the simple genetic algorithm*

**Keywords:** *Genetic Algorithm, Association Rule Learning, Stock Selection, Cross Over Operator, Mutation, Root Mean Square value.*

## 1. Introduction

The performance of Standard Genetic Algorithm (SGA) [1] is enriched by adapting the linkage learning methods [8] inside the Genetic Algorithms which learn the relationships among genes in the chromosomes. The success of a SGA relies upon a good coding scheme that puts genes belonging to the same building blocks together. This makes it easier for a Genetic Algorithm to traverse the fitness landscape towards an optimal solution.

Linkage problem [10] [18] is an ordering problem of the chromosome and it is addressed to the issue of building-blocks (BBs) identification or linkage learning.

To identify the best building blocks in a chromosome, techniques like Bayesian Optimization Algorithm (BOA) [21], Estimation Distribution Algorithm (EDA) and Gene Expression Messy GA (GEMGA) is used. Most of these techniques follow random approaches or probabilistic approaches. Most of the linkage learning techniques do not have standard procedure or algorithms to follow in order to identify the linkages in chromosomes. These approaches may take more computation time and are also difficult if the representation is too complicated. So there is a need for improvisation in linkage learning techniques.

This paper proposes a new linkage learning method known as Association Rule Linkage Learning (ARLLO) is a multi metric one, which utilizes one or more user-defined measurements to determine the solution quality. The proposed linkage learning technique finds relationship between input variables. The proposed association rule linkage learning technique uses “if then else rule” based approach [13] [17] to find the linkages that are existing in chromosomes. The proposed method has been employed in the Indian Stock Market to identify the linkages between the selected financial indicators and find out the performance of the stocks which are worthy of investments.

## 2. Steps Involved In Proposed ARLLO Method

```
Step 1: Set the population size (N);
Step 2: Generate initial population;
Step 3: fi = fitness improvement;
Step 4: Support_value = 0;
Step 5: Evaluate fitness of all individuals (N);
for all individuals (N)
Step 6: Check fitness improvement;
Step 7: if (fi exists in a candidate chromosome)
        then (identify BBs in a candidate
              chromosome and store in an array);
              increment Support_value by 1;
              put candidate chromosome in a higher
              group
        else
              Support_value = 0;
              put candidate chromosome in a lower
              group
        end if
Step 8: Repeat step 6 and step 7 for all individuals in the population
Step 9: Apply Multi Population GA (MPGA)
Step 10: Repeat from step 5 to 9 until the convergence is met.
```

Fig. 1: Algorithm of ARLLO in SGA

In several studies on the stock market, the problem of identifying a good stock has been solved by Hidden Markov [15] Model, Artificial Neural Network, Simulated Annealing [4] and Simple Genetic Algorithm [2]. These heuristic algorithms for the stock selection problem increase the computation overhead. To reduce the computation overhead, and to prove the efficiency of ARLLO [14], a stock selection problem has been chosen as a case study problem. The proposed ARLLO can also be used for prediction based problems which uses past data for prediction. Hence the ARLLO predicts the stocks and advise the investor to invest in the particular stock.

As per the algorithm given in figure1, the fitness of each individual is checked for improvement. If the fitness improvement is there, then it is assumed that the BBs exist in the chromosome. Building blocks are important in a chromosome to obtain the optimal solutions. If BBs exists in a chromosome, then the linkage is said to be good and it also improves the fitness values. Those BBs (stronger genes) have been identified and stored in the array. The ARLLO continues to find and identifies BB in all chromosomes in the population. If a chromosome has three BBs (three substrings) or three BB (three single bits referred as strong genes) in a chromosome then the Support\_value is three. Based on the Support\_value the ARLLO method divides the population into two groups' namely higher population group and lower population group. Higher the Support\_value of any chromosome can be given to the higher group population otherwise the chromosome is given to the lower group population which is shown in steps 6 and 7 of figure1.

The steps 6 and 7 have been applied to all individuals in the population. According to their Support\_value, the individuals are transferred to their respective groups.

Once the population has been divided into groups, apply Multi Population GA (MGA) to find the optimal stocks among the list of companies stocks. Steps 5 to 10 are iterated until the convergence is met.

## 3. Financial indicators used for stock selection

For the purpose of ranking the stocks the study uses the financial indicators such as Price to Earnings ratio, Book Value per Share, Current Ratio, Earnings per Share and Price to Book Value ratio. Along with these ratios Annual Price Return is also used. These ratios and annual price return are calculated [5] [14] to find the better performing stocks among the given stocks using the proposed algorithm.

## 4. GA using ARllo for stock selection process

The algorithm given figure1 is common for any optimization problem. The Genetic Algorithm with ARLLO approach for stock selection process is explained below with respect to the steps given in the algorithm (Figure1).

The step-wise application of this algorithm to the stock selection problem is adopted as step 1, step 2, step 3, step 4 (variable declaration are given), step 7, step 8, step 5, step 6, step 9 and step 10 and it is explained below:.

Step 1: To solve any problem algorithmically, the input variables have to be efficiently defined. Once the input parameters are defined, it should be properly encoded or represented known as chromosome representation which is shown in figure 2. In the present study of stock selection, each chromosome is composed of five input parameters, namely Price to Earnings ratio (P/E), Book Value per Share (BV), Current Ratio, (CR), Earnings per Share (EPS) and Price to Book Value ratio (P/B).

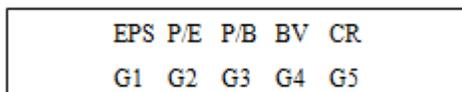


Fig. 2: Chromosome Representation of Financial Ratio Indicators

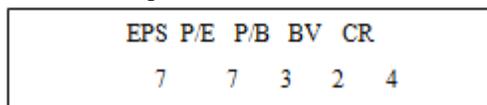


Fig. 3: Status Values (Genes) of Financial Ratio Indicators

For each input variable (financial ratio indicators), design eight statuses representing different qualities in terms of different range of values varying from 0 (extremely poor) to 7 (extremely good). This status information is used as gene values for the five financial ratio indicators. For example, in figure 3, the genetic value of EPS is ‘7’. It means the range of EPS value for a particular company falls in status ‘7’. If the status is ‘7’ then the financial indicator EPS is high for that particular company and it may increase the company’s annual return. The remaining individuals in the population are encoded and used in a similar way as shown in figure 3.

For the purpose of representing the eight statuses of a financial indicator, the values of the financial ratio indicators of the selected companies are grouped into eight classes. The grouping has been done by following the procedure as follows:

The difference between the maximum value and minimum value of the selected companies is divided by eight. The resultant quotient is taken rounded off to the nearest ten values and this value has been taken as ‘k’ is referred as class interval. The values of the financial ratio indicators less than ‘k’ (class interval) are assigned as the status of zero. More than ‘k’ value but less than ‘2k’ is assigned as the status of 1. Greater than ‘2k’ but less than equal to ‘3k’ is the status of ‘2’. Likewise the statuses are assigned till 6th statuses. The values of financial ratio indicators which are greater than ‘7k’ are assigned the status ‘7’. For the purpose of providing an example, the segregation of the statuses of EPS for the selected companies is presented in TABLE I:

The segregation of statuses for the remaining financial ratio indicators has been carried out in the same manner as that of EPS.

TABLE I: Encoding Of Eps Indicator

Sl. No.	EPS Value	Status
1	> 180	0
2	180 – 360	1
3	360 – 540	2
4	540 - 720	3
5	720 - 900	4
6	900 - 1080	5
7	1080 - 1260	6
8	< 1260	7

Step 7 & 8: Once the chromosome (financial ratio indicators) encoding is done, the proposed ARLLO check the status of the most influencing variable EPS and P/E ratio. If the status of these financial ratio indicators is above ‘3’ then these chromosomes (financial ratio indicators) are given to the elite/higher group population (PH) else the chromosomes (financial ratio indicators) are given to the lower group population (PL). In a lower group population the Support\_value of FRI is ‘0’. The segregation of chromosomes to their respective groups according to their status is shown in the following TABLE II.

TABLE II: Population Groups and Their Statuses

EPS	P/E	P/B	BV	CR	P <sub>G</sub>
0 - 3	0 - 3	0 - 3	0 - 3	0 - 3	P <sub>L</sub>
4 - 7	4 - 7	4 - 7	4 - 7	4 - 7	P <sub>H</sub>

Where PG = Population Group,

PL = Lower Population Group,

PH = Higher Population Group.

After dividing the initial population into groups, the ARLLO analyzes the relationship between stock return and FRI. Based on the relationship, the stocks are ranked. The algorithm assigns the ranks for the socks using the hidden relationship measured by it using the past data of APR and its association with the selected financial ratio indicators which is shown below.

In stock selection, the input variables are defined as

$N = \{EPS, P/E, BVPS, P/B, CR\}$  and their association rule is

$$\{EPS, P/E, BVPS, P/B, CR\} \rightleftarrows \{APR\}$$

The above equation indicates the relationship between the financial ratio indicators such as EPS, P/E, BV, P/B, CR and the APR. The APR value is dependent value and it is associated with the financial ratio indicators kept on the left hand side of the rule. For example, higher the financial ratio indicators values increase company's performance in terms of annual return.

According to the step 7, the proposed association rule based linkage learning process checks the statuses of each financial indicator (gene). If the status of each financial indicator is 7 then the Support\_value of this particular chromosome is 5 i.e. all five financial ratio indicators (five genes) are having the status '7'. The financial ratio indicators EPS and P/E are the most important variable that influences the annual price return when compared to the other financial ratio indicators. The higher status of these financial ratio indicators may provide good returns and these variables are referred as building blocks. The chromosome having BBs is said to be a strong individual and the ARLLO put this individual in the higher group population. The same process is applied for all individuals in the population (step 8) and builds two groups of population.

Once the algorithm has explored and identified the hidden relationship between APR and the financial ratio indicators that are existing in the population groups it assigns the rank thereafter for the stocks of the companies using only the financial ratio indicators. After the ranks are assigned by the ARLLO, it is validated by the actual rank. The Root Mean Square Error (RMSE) is used as a fitness function to validate the output of the ARLLO which is described below.

#### 4.1. Fitness Function

Step 5 & 6: Thus, the fitness function can be designed to minimize the root mean square error of the difference between the financial indicator derived ranking and the next year's actual ranking of all the listed companies for a particular chromosome, representing by.

$$RMSE = \sqrt{\frac{1}{m} \sum_{t=1}^m (R_{derived} - R_{actual})^2}$$

The fitness function in the present study is prepared using RMSE for which the formula is presented above. The idea in calculating RMSE is to measure the error in the estimation by making the comparison between the actual ranks assigned using the APR and the ranks assigned by the algorithm for all 100 companies (population size).

The RMSE is calculated as the mean of the sum of the squared differences between the derived ranks by the algorithm and the actual ranks assigned by calculating the APR values of the stocks in the given period. The smaller the value of RMSE better is the performance of the proposed algorithm. After validating the results of ARLLO for the first generation, the population for the next generation is explored by the multi population GA which is explained in step 9.

Step 9: According to the step 9 in the algorithm (figure 1) multi population GA has been formulated by the proposed ARLLO operator. The better chromosomes with strong genes (financial ratio indicators) are selected using roulette wheel selection method from each group to build strong individuals for the next generation population.

Following the selection mechanism the multi parent crossover operator is applied and they generated offspring for the next generation. Recombinant operators (crossover operators) engage in providing offspring by combining two highly fit parents. It changes the fitness value of the offspring from that of the parents and thus helps in the diversity of the Genetic Algorithm. It is also used to explore different regions of the search space and thus aids in the exploration policy of Genetic Algorithm. The illustration of multi parent crossover operation is shown in figure 4.

The multi parent crossover operation [12] [19] randomly takes three parents, namely parent 1 (P1) and parent 2 (P2) and parent 3 (P3). It compares each allele of parent 1 (P1) and parent 2 (P2). The common gene of both the parent is given to the offspring1 (O1). For example the first allele of P1 and P2 is '7' so it is given to the offspring1 (O1). The next allele of both the parent (P1 and P2) is different that is '6 & 7'. So it copies the second allele of the third parent (P3) to O1. This process is continued for all the genes in a chromosome and produces the offspring1 (O1).

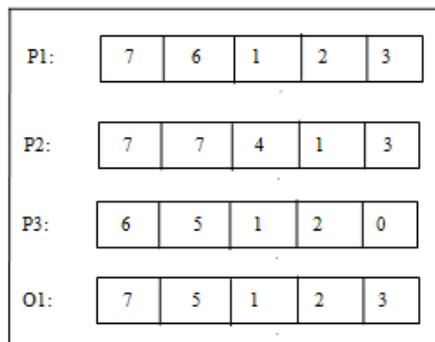


Fig. 4: Status of FRI in Crossover Operation

After the crossover operation, the chromosomes are mutated using the swap mutation [16] [20] technique with a probability of 0.05 is shown in figure 6.7. The chromosome (P1) is randomly chosen from any of the group for mutation. The swap mutation randomly selects two gene positions and their values (status of two financial ratio indicators) are exchanged to produce a new chromosome (O1) (financial ratio indicators).

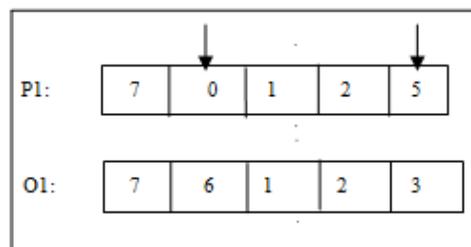


Fig. 5: Status of FRI in Mutation

The obtained offspring from selection technique, crossover operation and mutation are kept in the next generation.

Step 10: The above ARLLO process is continued till the convergence is met. That is, the final generation will be judged once the optimized result is obtained.

## 5. Experimental Results and Performance Comparison

The data for the experimental analysis has been collected from “PROWES”, a database of Centre for Monitoring Indian Economy, India. The sample data has been collected for a period April 2010 to March 2013. This data has been collected from 100 companies randomly selected from the list of 200 companies used for calculating S&P BSE 200 Index. This index is a broad based index used to represent the market movements in the Bombay Stock Exchange (BSE).

In Genetic Algorithm optimization [22] the population size is finite, which influences the sampling ability of a Genetic Algorithm. For the experimental purpose, the data for 100 companies for three years has been considered.

The Genetic Algorithm parameter settings for the experimentation are shown in TABLE III.

TABLE III: Ga Parameter Settings

Population Size	100
Selection	Roulette Wheel Method
Selection Rate	10 %
Mutation	Swap Mutation
Mutation Rate	5%
Crossover	Two Point Crossover

The APR values of the selected companies are calculated [14] and based on the APR values, the stocks are ranked and named as Ractual. Based on the past financial ratio indicator values and the APR values, the SGA computed ranks is referred as Rsga. Similarly the ranks computed by the ARLLO are Rarllo. Using the Rsga and Ractual, root mean square error has been calculated [14] for the results of SGA. Similarly using Rarllo and Ractual,, the root mean square error has been calculated for the results of ARLLO. These results are given in the TABLE IV for the three financial years 2010 – 2011 to 2012-2013. Based on the results of ARLLO the investors could invest their shares in top most ranked companies which are likely to have good performance.

TABLE IV  
Fitness Values of  $R_{sga}$  And  $R_{arllo}$

Fitness value/ Algorithms	$R_{sga}$	$R_{arllo}$
RMSE (2010 - 2011)	12.3288	5
RMSE (2011-2012)	13.3041	4.6904
RMSE (2012 - 2013)	10.4403	6.7823

The performance comparison of the proposed method with simple genetic algorithm is shown in the figure 6 given below.

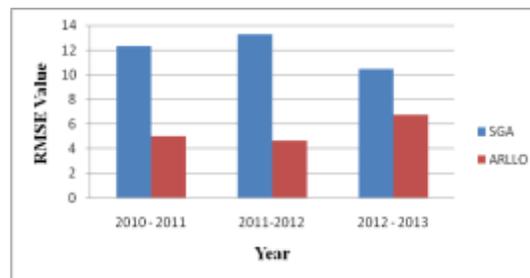


Fig. 6: RMSE Comparison of ARLLO with SGA

The smaller the value of RMSE better is the performance of the proposed algorithm. The RMSE value of the proposed ARLLO is less for all the three years than the SGA. From the TABLE IV, it is concluded that the proposed ARLLO operator found to perform better when compared to the SGA.

## 6. Conclusion

In this research, the proposed Association Rule Linkage Learning Technique is used to identify the best performing stocks in the market has been used. This has been proved in the real-time data collected from the Indian popular and very old Stock Market, Bombay Stock Exchange. Thus, it is concluded that the performance of the proposed Linkage Learning technique ARLLO is significant when compared to the performance of the SGA.

## 7. Acknowledgment

The suggestions and motivation of Dr. K.Vivekanandan, Professor of Computer Science, Pondicherry Engineering College, Puducherry, India and Dr.S.Siva Sathya, Associate Professor, Department of Computer Science and Engineering, Pondicherry University, Puducherry, India are gratefully acknowledged.

## 8. References

- [1] Beasley .D, Bull, and Martin .R. (1993). An Overview of genetic algorithms: Part 1, Fundamentals, University Computing, vol. 2, pp. 58-69, 1993.

- [2] R.Lakshmi, K.Vivekanandan and R.Brintha, A New Biological Operator in Genetic Algorithm for Class Scheduling Problem, *International Journal of Computer Applications (IJCA)*, Volume 60, Issue 12, December 2012, ISSN: 0975 ñ 887.
- [3] Ayed A. Salman, Kishan Mehrotra, and Chilukuri K. Mohan (2000). Adaptive linkage crossover in evolutionary computation. *Proceedings of Evolutionary Computation*, Vol.8, Issue 1(3).
- [4] Booker. (1987). Improving search in genetic algorithms, *Genetic Algorithms and Simulated Annealing*. Pitman, chapter 5, pp 61-73.
- [5] Chengxiong Zhou and et al. (2006). Selecting Valuable Stock Using Genetic Algorithm. Springer-Verlag Berlin Heidelberg pp. 688–694.  
[http://dx.doi.org/10.1007/11903697\\_87](http://dx.doi.org/10.1007/11903697_87)
- [6] Chen YP, Goldberg D.E. (2004). Convergence time for the linkage learning genetic algorithm, in *Proceedings of IEEE Congress on Evolutionary Computation*, pp. 39–46.
- [7] R. Lakshmi and K. Vivekanandan, Heuristics Based Learning on Human Psychology, *International Journal of Computer Applications*, Feb 2013, ISSN: 0975 ñ 8887.
- [8] R. Lakshmi and K. Vivekanandan, Interference Induced Silencing in Travelling Salesperson Problem using Linkage Learning Genetic Algorithm, *International Journal of Engineering & Science and Research (IJESR)*, Volume 3, Issue 3.
- [9] Chuang, C.-Y., & Chen, Y.-p. (2007). Linkage identification by perturbation and decision tree induction. *Proceedings of IEEE Congress on Evolutionary Computation (CEC 2007)*, pp. 357–363.  
<http://dx.doi.org/10.1109/CEC.2007.4424493>
- [10] R. Lakshmi and K. Vivekanandan, Performance Analysis of Linkage Learning Techniques in Genetic Algorithms, *International Journal of Research in Engineering and Technology*, Volume 02, Issue: 12, Dec 2013, EISSN: 2319-1163, PISSN: 2321-7308.
- [11] David and et al. (2007). Higher-Order Linkage Learning in the ECGA. *ACM*.
- [12] R. Lakshmi and K. Vivekanandan, A Novel Methodology for Genetic Algorithms in Crossover Operation: Segment Replacement Operator, *International Journal of Innovative Research & Development*, Volume 2, Issue 14, Jan 2014, ISSN: 2278- 0211.
- [13] De Jong, K. A. & Spears, W. M. (1991). Learning concept classification rules using genetic algorithms. In *Proceedings of the Twelfth International Conference on Artificial Intelligence (IJCAI-91)*, pp. 651-656.
- [14] R. Lakshmi, K. Vivekanandan and S.Amilan, New Linkage Learning Technique in Genetic Algorithm for Stock Selection Problem, *International Journal of Advanced Research in Computer Science (IJARCS)*, Vol. 5, Issue 2, March 2014, ISSN: 0976 ñ 5697.
- [15] Goldberg, D.E and Segrest, P. (1987). Finite Markov chain analysis of genetic algorithms. *Proceedings of the the Second International Conference on Genetic Algorithms*, pp 1-8.
- [16] R. Lakshmi and K. Vivekanandan, A Novel Hybrid Crossover Operator for Genetic Algorithm, *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*, Volume 4, Issue 4, March 2014, ISSN: 2277 128X.
- [17] Goldberg, D.E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison Wesley Longman, Inc., ISBN: 0-201-15767-5.
- [18] R. Lakshmi and K. Vivekanandan, Gene Silencing in Linkage Learning GA, *National Conference on Future Computing (NCFC 2012)*, Pondicherry University, India.
- [19] R. Lakshmi and K. Vivekanandan, “An Analysis of Recombination Operator in Genetic Algorithms, *IEEE International Conference on Advance Computing (ICoAC) 2013*, MIT, Chennai, India (Awarded Best Paper).  
<http://dx.doi.org/10.1109/ICoAC.2013.6921954>
- [20] R.Lakshmi and K.Vivekanandan, Performance Analysis of a Novel Crossover Technique on Permutation Encoded Genetic Algorithms, *IEEE International Conference on Advance Engineering Technology (ICAET) 2014*, EGS Pillai Engineering College, Nagapattinam, India.
- [21] Larranaga, P., Kuijpers, C.M.H., Poza, M., y Murga, R.H. (2009). Decomposing Baysian Networks: Triangulation of the Moral Graph with genetic Algorithms. *Statistics and Computing*.
- [22] R.Lakshmi and K.Vivekanandan, Performance Evaluation of a new Crossover on Permuted Genetic Algorithm, *International Conference on Recent Trends in Engineering and Technology (ICRTET) 2014*, Mount Zion College of Engineering and Technology, Pudukottai, India.