Optimization of Fragment based Mining through Genetic Algorithm

Rajesh V. Argiddi
Assit.Prof.Department of computer Science & Engginering, Walchand Institute of Technology,
Solapur, India

S .S. Apte, Ph.D
Head of Computer Science & Engginering Department, Walchand Institute of Technology,
Solapur, India

ABSTRACT

The approach stated in this paper mainly focuses on generating optimized rules in fragment based association mining using genetic algorithm. we call this approach as Genetic based Fragment Rule Mining. we designed a novel method for generation of optimized rule. In which a Fragment mining is used to generate the rules on which we use the optimization mechanism. This deals mainly with reducing the time and space complexity required in processing the data using fragment mining & generate strong rules using genetic algorithm. The results reported in this paper are very promising since the discovered rules are of optimized rules.

Keywords

Association Rule, Fragment Mining ,Stock Data , Genetic Algorithm.

1. INTRODUCTION

Data mining is one of the research field in the current corporate database and information decision making systems. From a technical point of view, it refers to the extracting previously unknown, the potentially useful patterns or knowledge from large databases including association rules, time series, classification, clustering, artificial intelligence, statistics etc. The term data mining has mostly been used by statisticians, data analysts, and the management information systems communities. The phrase knowledge discovery in databases (KDD) refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in KDD process. In this paper data mining techniques are applied on stock market data in order to predict the stock scenario.

Financial institutions such as stock markets produce huge datasets that build a foundation for approaching these enormously complex and dynamic problems with data mining tools. Association rule [9] is a technique to detect the hidden facts in large dataset and draw interferences on how subsets of items influence the presence of other subsets. Association mining mostly fits best to direct customer oriented businesses, the rules generated from this technique is suitable for gaining knowledge .Association rule mining aims to find strong relation between attributes. Apriori algorithm best for single attribute rule generation, Apriori requires more time as the number of transactions gets increased. After this FITI (First Intra-transaction then Inter-transaction) algorithm was introduced, but the drawback of FITI is its efficiency decreases as the number of transactions increases. To overcome this WangZhong [1] proposed a technique Granule based mining, which allows group of transactions based on common features of the transactions. Further, Prof.R.V.Argiddi [2] had used this approach granule based mining as fragment based mining ,in which they elaborated the work by evaluating a single attribute behavior based on group of attributes and also validated their predictions.

The original problem addressed by fragment rule mining was to find a correlation among sales of different products from the analysis of a large set of data & generates large number of rules. But users are not interested in all association rules, they are just concerned about the associations among condition attributes and decision attributes. To overcome this we introduced genetic approach for optimization of rules generated from fragment based mining [3]. Genetic algorithm is a family of computational models based on principles of evolution and natural selection .These algorithms convert the problem in a specific domain into a model by using a chromosome-like data structure and evolve the chromosomes using selection, recombination, and mutation operators. The range of the applications that can make use of genetic algorithm is quite broad [10].

2. BACKGROUND

2.1 Association Rule

One of the most important research areas in the field of Data mining is ARM. Association rules are used to identify relationships among a set of items in a transactional dataset. In the previous research, The problem of discovering association rules was first introduced in [5] and an algorithm called AIS was proposed for mining association rules. For last fifteen years many algorithms for rule mining have been proposed.

The task of mining association rules over market basket data is considered a core knowledge discovery activity. Association rule mining provides a useful method for discovering relations among items belonging to customer transactions in a market basket database. Covering sets are important because they are the building blocks to obtain association rules with a given confidence and support. Essentially, association mining is about discovering a set of rules that is shared among a large percentage of the data. Association rules mining tend to produce a large number of rules. The aim is to find the rules that are useful to users.

2.2 Fragment Rule Mining

Wanzhong Yang has proposed one innovative technique to process the stock data named Granule mining technique, which reduces the width of the transaction data and generates the association rules [1].Further, Prof.R.V.Argiddi has proposed fragment based mining, which deals mainly with reducing the time and space complexity involved in processing the data in association rule mining technique[2].

As in granule mining, fragment based approach fragments the data sets into fragments for processing thereby reducing the input size of data sets fed to the algorithm. Fragment based mining basically deals in grouping the stocks of different IT companies shares and producing some sort of association among the shares , which helps in predicting the IT market. It groups the shares based on small scale companies market data as condition attribute & large scale companies market data as decision attribute .

In contrast to granule mining, in fragment based mining the covering set of condition attribute are formed and decision attributes are aggregated for obtaining generalized association rules.

2.3 Strengths and weaknesses of fragment based rules

The first and maybe, one of the biggest advantages of association rules is that the results produced are very easy to be understood. Furthermore, it is a technique that can be used when we do not really know where to start from but they want to examine the trends in market. Most probably it is the best technique to use when they do not have something particular in mind; when they do not want to predict the value of a specific attribute. The computations that need to be performed are not very complex; that is another strong point of the method.

On the other hand, it has drawbacks associated with it. Despite the fact that the computations involved are simple, the search space can be very large with the presence of a relatively small number of items , in fact the relationships between the number of items and the search space is of exponential magnitude. The implementation of the algorithms tends to lead to a very large number of association rules. In most of the case is that they will be confusing and we will have to examine each one of them very carefully to determine if it is of any interest to us.

Kannika Nirai Vaani M,E Ramaraj has proposed new approach to generate association rules i.e promoting the faster generation of frequent item sets, so that to offer useful rules in an effective and optimized manner with the help of Genetic Algorithm[4][5].

3. METHODOLOGY

Genetic based fragment rule mining is two phase approach, while in first phase of this method we perform Fragment rule mining which is based on association rule & granule mining. & second phase is of optimization of results using genetic algorithm. The Detailed design of this approach is shown in Fig 1 as below.

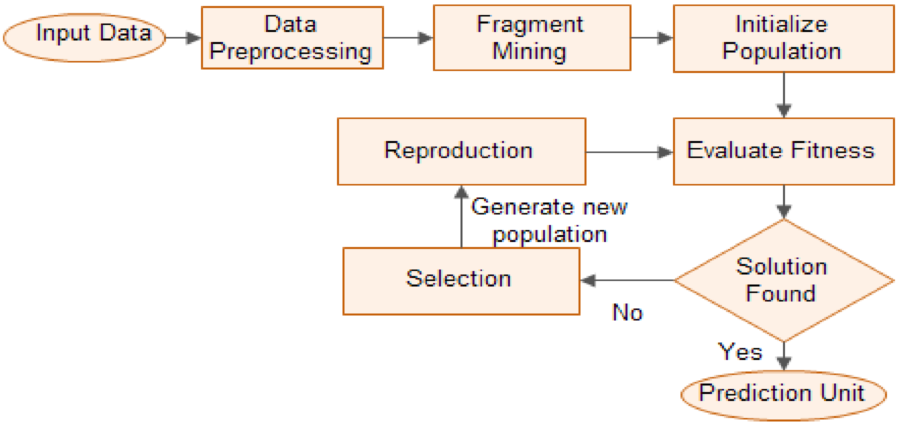
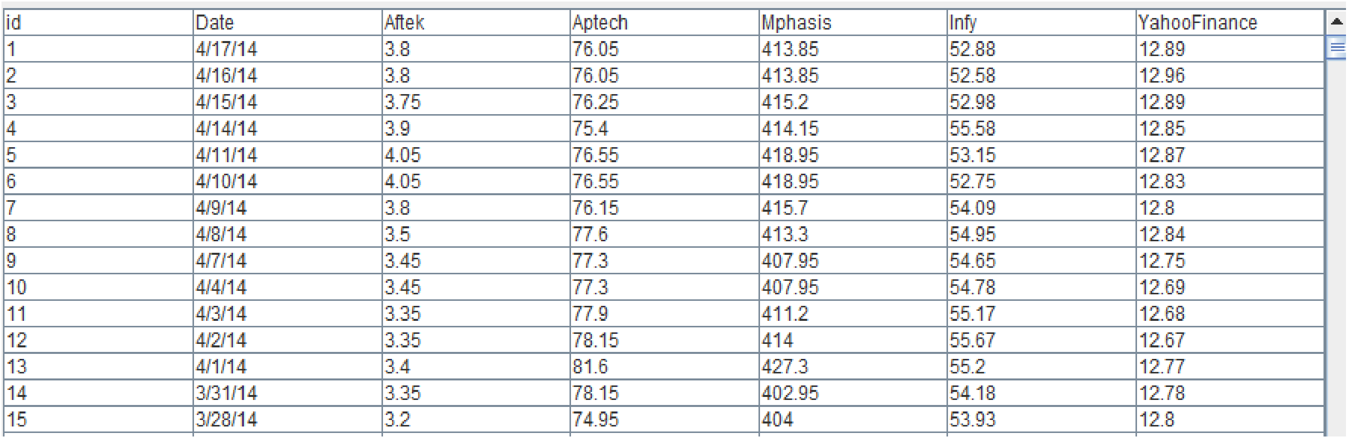


Figure 1.Data flow diagram

3.1 Fragment based Mining

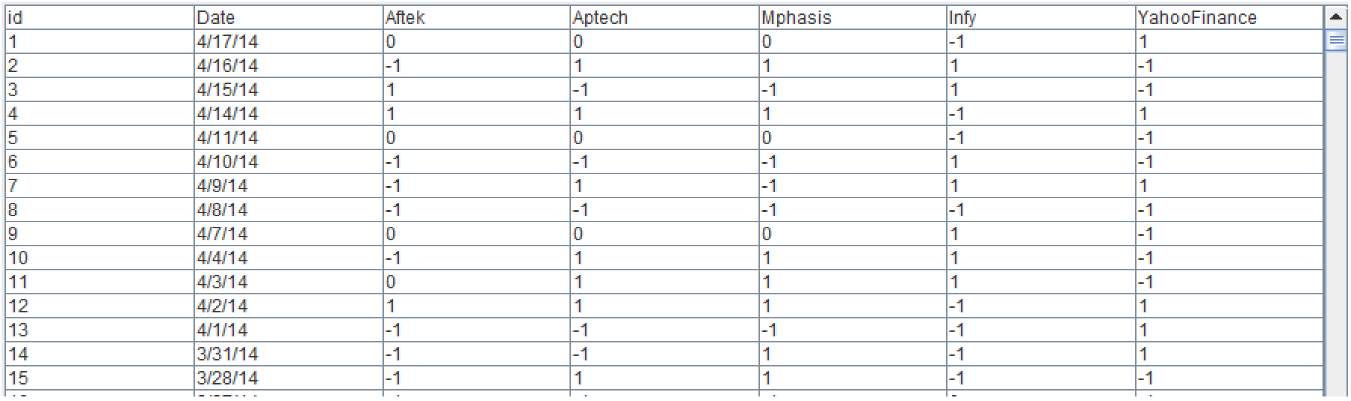
In this approach we first differentiate the companies based on small and large scale, grouping is done based on the capitalization of the company. Then we select some of the companies from the small scale and some from the large scale. This files are stored in the form of excel sheet at the back end. Once we select the companies, data is extracted from the excel sheet and we can apply the fragment mining on this data and generate the rules for prediction. Before generation of rules it gives frequent pattern set. i.e covering set . We consider only the date and price of all the companies as shown in Fig2.



id	Date	Aftek	Aptech	Mphasis	Infy	YahooFinance
1	4/17/14	3.8	76.05	413.85	52.88	12.89
2	4/16/14	3.8	76.05	413.85	52.58	12.96
3	4/15/14	3.75	76.25	415.2	52.98	12.89
4	4/14/14	3.9	75.4	414.15	55.58	12.85
5	4/11/14	4.05	76.55	418.95	53.15	12.87
6	4/10/14	4.05	76.55	418.95	52.75	12.83
7	4/9/14	3.8	76.15	415.7	54.09	12.8
8	4/8/14	3.5	77.6	413.3	54.95	12.84
9	4/7/14	3.45	77.3	407.95	54.65	12.75
10	4/4/14	3.45	77.3	407.95	54.78	12.69
11	4/3/14	3.35	77.9	411.2	55.17	12.68
12	4/2/14	3.35	78.15	414	55.67	12.67
13	4/1/14	3.4	81.6	427.3	55.2	12.77
14	3/31/14	3.35	78.15	402.95	54.18	12.78
15	3/28/14	3.2	74.95	404	53.93	12.8

Figure2.Selectedcompanylist

Now we convert the data into 1's and 0's, this is done by performing operation such as Transaction1=Transaction2 – Transaction1, and if transaction1>0 we put 1, if transaction1<0 we put -1, otherwise 0, so in this manner we build the following Fig 3.



id	Date	Aftek	Aptech	Mphasis	Infy	YahooFinance
1	4/17/14	0	0	0	-1	1
2	4/16/14	-1	1	1	1	-1
3	4/15/14	1	-1	-1	1	-1
4	4/14/14	1	1	1	-1	1
5	4/11/14	0	0	0	-1	-1
6	4/10/14	-1	-1	-1	1	-1
7	4/9/14	-1	1	-1	1	1
8	4/8/14	-1	-1	-1	-1	-1
9	4/7/14	0	0	0	1	-1
10	4/4/14	-1	1	1	1	-1
11	4/3/14	0	1	1	1	-1
12	4/2/14	1	1	1	-1	1
13	4/1/14	-1	-1	-1	-1	1
14	3/31/14	-1	-1	1	-1	1
15	3/28/14	-1	1	1	-1	-1

Figure 3. Converted table

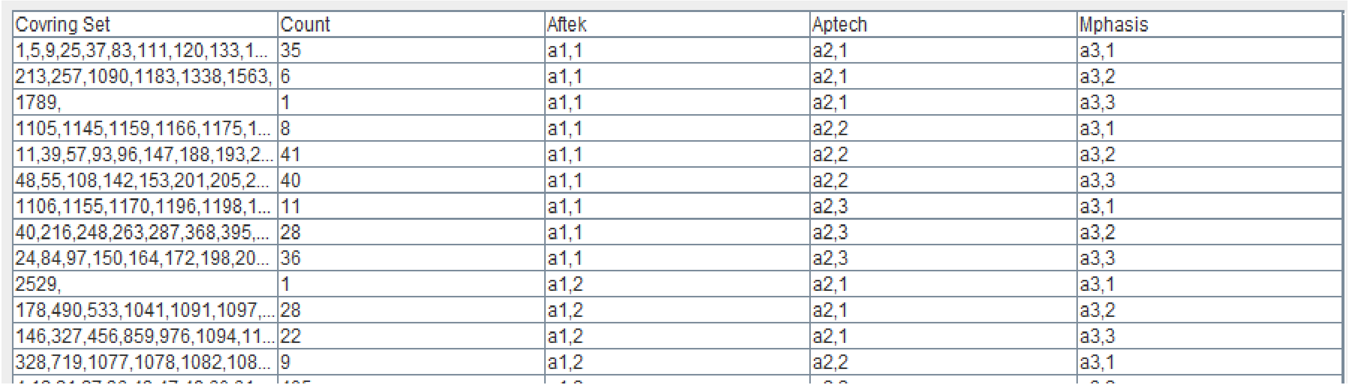
Fragment based mining is based on two tier fragmentation, condition fragments and decision fragments, That's why we divide dataset as small scale & Large scale companies. i.e Decision of large scale company depends on condition of small scale companies .

small scale (Condition Attribute) : Aftek, Aptech, Mphasis (a1,a2,a3)

.

Large scale (Decision Attribute): YahooFinance, Infy (b1,b2)

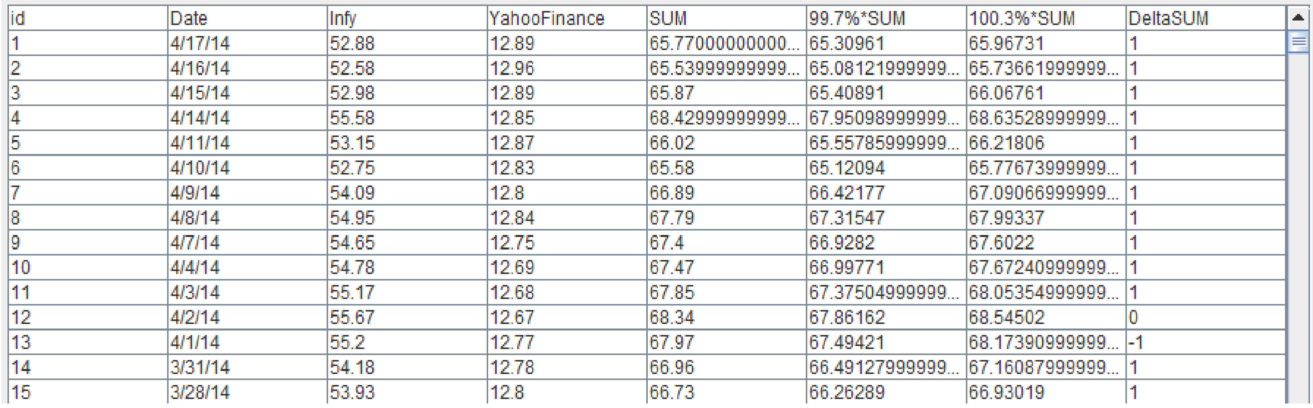
After that group the transactions of condition granules based on similar rows and form the covering set



Covring Set	Count	Aftek	Aptech	Mphasis
1,5,9,25,37,83,111,120,133,1...	35	a1,1	a2,1	a3,1
213,257,1090,1183,1338,1563,	6	a1,1	a2,1	a3,2
1789,	1	a1,1	a2,1	a3,3
1105,1145,1159,1166,1175,1...	8	a1,1	a2,2	a3,1
11,39,57,93,96,147,188,193,2...	41	a1,1	a2,2	a3,2
48,55,108,142,153,201,205,2...	40	a1,1	a2,2	a3,3
1106,1155,1170,1196,1198,1...	11	a1,1	a2,3	a3,1
40,216,248,263,287,368,395,...	28	a1,1	a2,3	a3,2
24,84,97,150,164,172,198,20...	36	a1,1	a2,3	a3,3
2529,	1	a1,2	a2,1	a3,1
178,490,533,1041,1091,1097,...	28	a1,2	a2,1	a3,2
146,327,456,859,976,1094,11...	22	a1,2	a2,1	a3,3
328,719,1077,1078,1082,108...	9	a1,2	a2,2	a3,1

Figure 4. Covering Set of conditional attribute

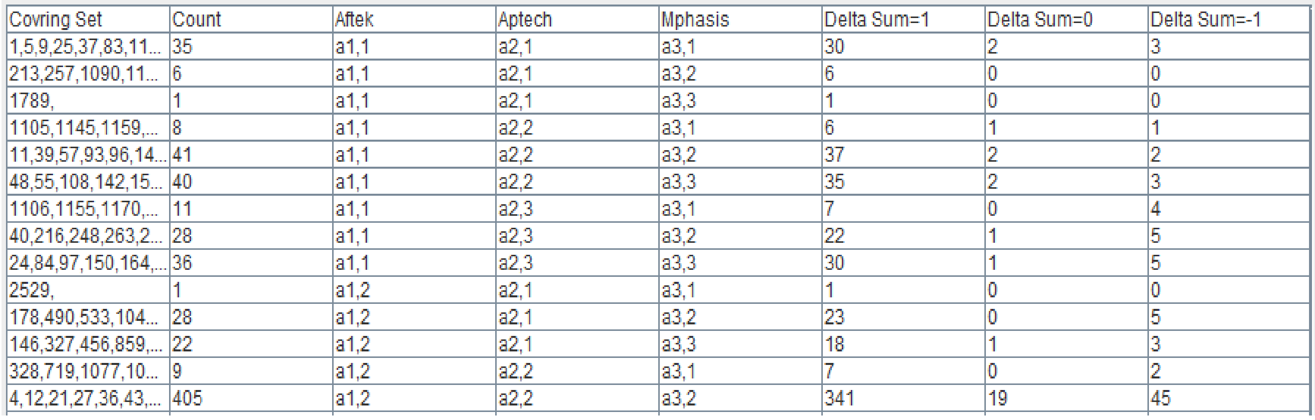
Next we consider the decision attributes; here we perform aggregation of all the Large scale companies and find the minimum and maximum range of these companies, we do this because the stock market is very fluctuating data. Further we will find the positive and negative gains i.e. evaluating 1 and 0 is done as below based upon the window size and processing the data in this window. Fig 5 Shows the aggregation of decision attributes .



id	Date	Infy	YahooFinance	SUM	99.7%*SUM	100.3%*SUM	DeltaSUM
1	4/17/14	52.88	12.89	65.77000000000...	65.30961	65.96731	1
2	4/16/14	52.58	12.96	65.53999999999...	65.08121999999...	65.73661999999...	1
3	4/15/14	52.98	12.89	65.87	65.40891	66.06761	1
4	4/14/14	55.58	12.85	68.42999999999...	67.95098999999...	68.63528999999...	1
5	4/11/14	53.15	12.87	66.02	65.55785999999...	66.21806	1
6	4/10/14	52.75	12.83	65.58	65.12094	65.77673999999...	1
7	4/9/14	54.09	12.8	66.89	66.42177	67.09066999999...	1
8	4/8/14	54.95	12.84	67.79	67.31547	67.99337	1
9	4/7/14	54.65	12.75	67.4	66.9282	67.6022	1
10	4/4/14	54.78	12.69	67.47	66.99771	67.67240999999...	1
11	4/3/14	55.17	12.68	67.85	67.37504999999...	68.05354999999...	1
12	4/2/14	55.67	12.67	68.34	67.86162	68.54502	0
13	4/1/14	55.2	12.77	67.97	67.49421	68.17390999999...	-1
14	3/31/14	54.18	12.78	66.96	66.49127999999...	67.16087999999...	1
15	3/28/14	53.93	12.8	66.73	66.26289	66.93019	1

Figure 5. Aggregation of large scale companies

By combining both the tables of large and small scale company we form the table, based on the covering set of the small scale companies we find the value of that id from the large scale company table, and group this based on positive gain(Delta sum=1), negative gain(Delta Sum=-1) or no change (delta sum=0) respectively. Finally after completing all the above procedure the rules will be generated as shown in Fig 6. and based upon this market behavior is been directly predicted on basis of confidence.



Covring Set	Count	Aftek	Aptech	Mphasis	Delta Sum=1	Delta Sum=0	Delta Sum=-1
1,5,9,25,37,83,11...	35	a1,1	a2,1	a3,1	30	2	3
213,257,1090,11...	6	a1,1	a2,1	a3,2	6	0	0
1789,	1	a1,1	a2,1	a3,3	1	0	0
1105,1145,1159,...	8	a1,1	a2,2	a3,1	6	1	1
11,39,57,93,96,14...	41	a1,1	a2,2	a3,2	37	2	2
48,55,108,142,15...	40	a1,1	a2,2	a3,3	35	2	3
1106,1155,1170,...	11	a1,1	a2,3	a3,1	7	0	4
40,216,248,263,2...	28	a1,1	a2,3	a3,2	22	1	5
24,84,97,150,164,...	36	a1,1	a2,3	a3,3	30	1	5
2529,	1	a1,2	a2,1	a3,1	1	0	0
178,490,533,104...	28	a1,2	a2,1	a3,2	23	0	5
146,327,456,859,...	22	a1,2	a2,1	a3,3	18	1	3
328,719,1077,10...	9	a1,2	a2,2	a3,1	7	0	2
4,12,21,27,36,43,...	405	a1,2	a2,2	a3,2	341	19	45

Figure 6. Fragment based mining

We got output as above, Now Consider,

Cg1 = transaction at id1 i.e a11,a21,a31,a41

For Cg1, dg1=1, dg2=0,dg3=-1 , N=2

Cg2 = transaction at id2 i.e a11,a21,a31,a43

For Cg1, dg1=0 , dg2=0,dg3=-1 , N=1

Cg3 = transaction at id3 i.e a11,a23,a33,a41

For Cg1, dg1=1 , dg2=0,dg3=-1, N=2

…………

Cgn = transaction at idn i.e a13,a23,a33,a43.

For Cgn, dg1=2 , dg2=0,dg3= 0, N=2

Now only ,These Cg & dg parameters are provided to genetic algorithm for optimization in form of linked list.

Cg1 { a11,a21,a31,a41} → Dg1 →Dg3

Cg3 { a11,a23,a33,a41} → Dg1 →Dg3

Cg6 { a13,a21,a31,a43} → Dg1

Cg8 { a13,a23,a33,a41} → Dg1 →Dg3

Cg9 { a13,a23,a33,a43} → Dg1

3.2 Rule Optimization by Genetic Algorithm

3.2.1 Data Encoding

Condition attributes covering set & aggregated sum of decision attribute are used as dataset. Cg is class array of number of covering sets of condition attribute , Each with status of small scale companies describes changes of 3 company shares (like a11,a21,a31 for Cg1) & 3 decision granules (Dg1,Dg2,Dg3) that describe possible gain of buying share after 3 days, based on small company share change. Genetic Algorithm directly not work on the raw data then whole data we have encoded in the form of Binary representation technique (0 and 1).

3.2.2 Fitness function

The most important part of Genetic Algorithm is a design of Fitness Function:

Let, Minimum confidence=50% ,Calculate Fitness function f(x),

f(x) =count of Dg(x)/ Total number of sum (x)

If (fitness function > min confidence)

Cg(x) → dg1 is rule otherwise, reject remaining rules.

3.2.3 Selection Strategy

The selection is done on the basis of individual fitness and confidence of individual whose fitness value is greater than its minimum confidence. Generate a new population by repeating following steps until the new population gives satisfied results.

3.2.4 Genetic Operation

The Genetic operators determine the search capability and convergence of the algorithm. Genetic operators hold the selection crossover and mutation on the population and generate the new population.

Selection**:** Select two parent chromosomes from a population on fitness basis ,better the fitness, the major chance to be selected.

Crossover**:** With a crossover probability cross over the parents to form offspring (children). If no crossover was performed, children are exact copy of their parents.

Mutation: Mutation probability mutate new offspring at each position in chromosome.

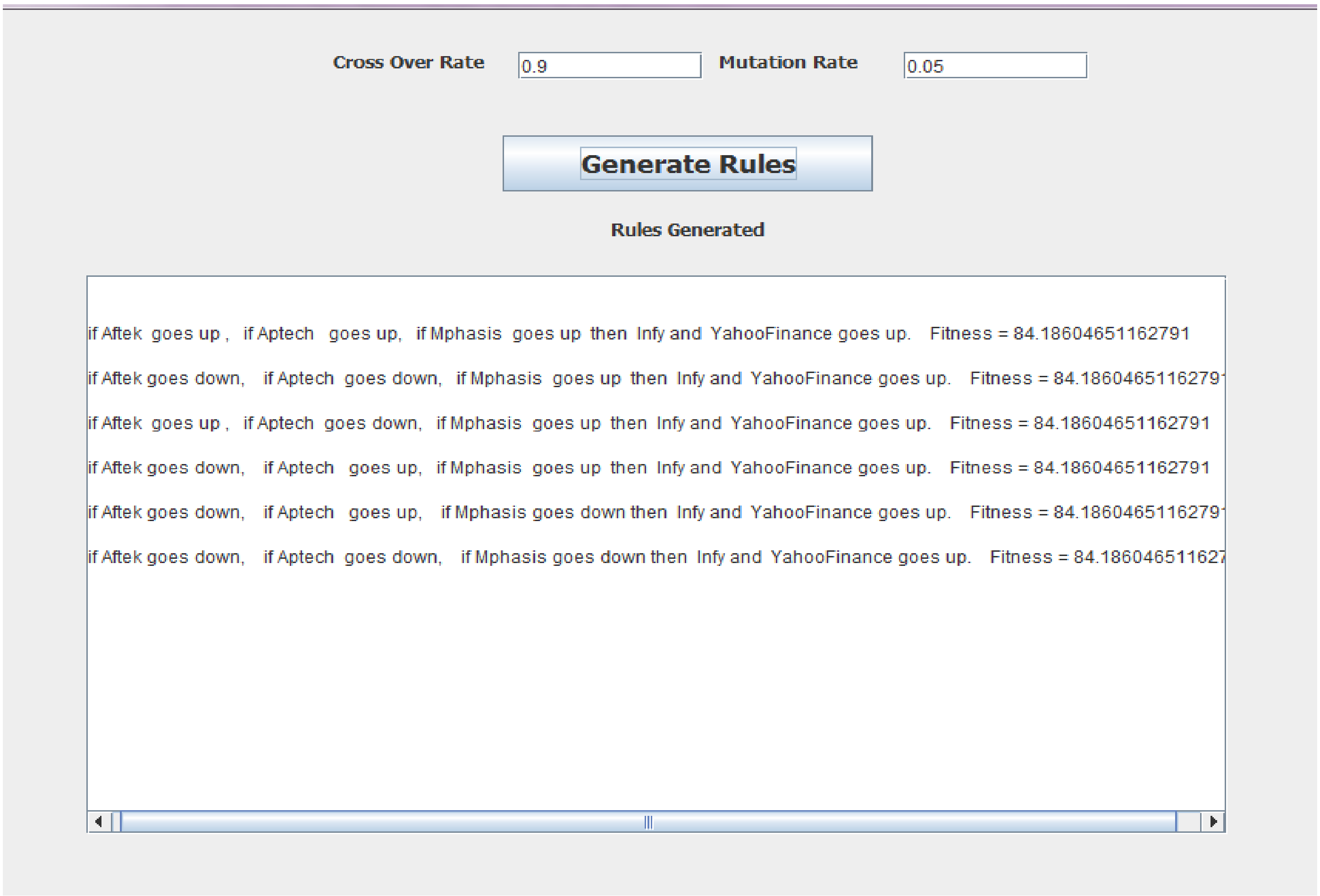
Accepting: Place new children (offspring) as a new population

3.2.5 Terminating condition

The algorithm finally extracts the rules that meet the confidence threshold given by users, so the final output is not one optimal rule, but rather a set of rules that meet the threshold.

i. 3.2.6 Rules extraction

The frequent rules are generated according to the fitness function and genetic operators. In order to mine the strong association rules finally, these rules must be extracted again. Extraction criteria are: output the rule which meets the minimum confidence given by users, otherwise abandon it.



Cross Over Rate 0.9 Mutation Rate 0.05

Rules Generated

```
if Aftek  goes up ,   if Aptech   goes up,   if Mphasis  goes up  then  Infy and  YahooFinance goes up.    Fitness = 84.18604651162791
if Aftek goes down,    if Aptech  goes down,   if Mphasis  goes up  then  Infy and  YahooFinance goes up.    Fitness = 84.1860465116279
if Aftek  goes up ,   if Aptech  goes down,   if Mphasis  goes up  then  Infy and  YahooFinance goes up.    Fitness = 84.18604651162791
if Aftek goes down,    if Aptech   goes up,   if Mphasis  goes up  then  Infy and  YahooFinance goes up.    Fitness = 84.18604651162791
if Aftek goes down,    if Aptech   goes up,    if Mphasis goes down then  Infy and  YahooFinance goes up.    Fitness = 84.1860465116279
if Aftek goes down,    if Aptech  goes down,    if Mphasis goes down then  Infy and  YahooFinance goes up.    Fitness = 84.186046511627
```

Figure 7. Genetic framework for optimization

4. EXPERIMENTS AND RESULTS

In this work, we have taken the original data sets of Bombay Stock Exchange (BSE) of different companies such as Infosys Aftek, Aptech, Mphasis etc from Yahoo Finance and try to find the association among the large IT companies and Small IT companies. This approach we are going to apply on real time Indian IT stock market data and find the efficiency of using this technique in producing the strong optimized rules for prediction. Here we have extracted last 10 years data i.e. from 01 March 2004 to 17 April 2014 and applied the Genetic based fragment rule mining approach and evaluate the rules for prediction. As explained in the methodology first we need to select some companies from both small and large scale companies. First, Perform Fragment rule mining as explained above & after that apply genetic in next phase. The setting of parameter: the size of evolutionary Generation=1000 population =100, crossover rate=0.5, mutation rate=0.08.

When we consider the above set then we get the following rule : Aftek (↓), Aptech (↓), Mphasis (↑) => Infosys, YahooFinance (↑) ……. Association Rule (1)

Fitness Value : 84.218 > Min Confidence

So from above rule we predict that when Aftek goes DOWN, Aptech goes DOWN, Mphasis goes UP and on certain day and a person investing after this observation and if he invests in INFY, Yahoo Finance will be in profit after 4 working days (as we are considering the window size as 4) Fig8. Shows Graphical view of the generated results of fragment mining & after optimization as it varies with minimum confidence.

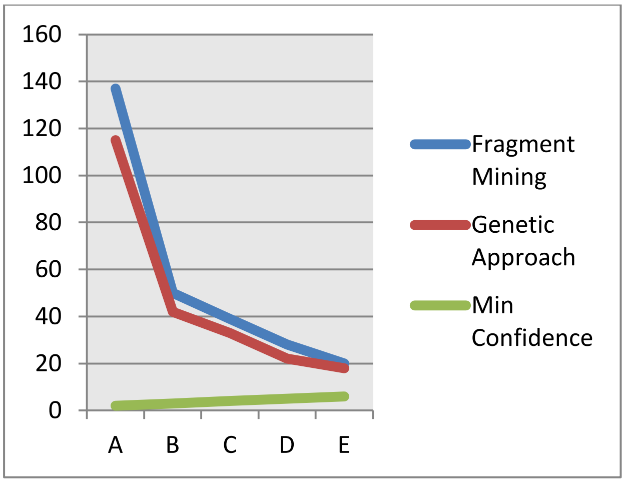


Figure.8 confidence vs. rule generation graph

5. CONCLUSION & FUTURE TREND

In this direction we optimize Fragment rule mining using strong fitness function of genetic algorithm. We conclude that this approach is an innovative & knowledge based, integrating a genetic algorithm to fragment based mining to obtain optimized rules that potentially be used for predictions in stock markets.

To make genetic algorithm more effective and efficient it can be incorporated with other techniques so it can provide a best Result

6. REFERENCES

[1]Wanzhong Yang, “Granule Based Knowledge Representation for Intra and Inter Transaction Association Mining”, Queensland University of Technology, July 2009

[2] R.V Argiddi,S.SApte ” study of association rule mining in fragmented item-sets for prediction of transactions outcome in stock trading systems” IJCET-2012

[3] Kannika Nirai Vaani M, E Ramaraj **“**An integrated approach to derive effective rules from association rule mining using genetic algorithm” IEEE2013 International Conference

[4] Kannika Nirai Vaani M, E Ramaraj” E-Rules: An Enhanced Approach to Derive Disjunctive and useful Rules from Association Rule Mining without Candidate Item Generation” IJCA-2013

[5] R. Agrawal, T. Imielinski, and A. Swami. “Mining association rules between sets of items in large databases”. In Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD '93), pages 207216, Washington, USA, May 1993.

[6] R.V Argiddi, S.SApte “ Future Trend Prediction of Indian IT Stock Market using Association Rule Mining of Transaction data” IJCA-2012.

[7] Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K. *“Optimized association rule mining using genetic algorithm”*, Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, 2009

[8] Prashant S. Chavan, Prof. Dr. Shrishail. T. Patil” Parameters for Stock Market Prediction” IJCTA | Mar-Apr 2013 Vol 4 (2),337-340

[9] Kalyanmoy Deb, “Introduction to Genetic Algorithms”, Kanpur Genetic Laboratory (Kangal), Depart of Mechanical Engineering, IIIT Kanpur 2005.

[10] Nikhil Jain,Vishal Sharma,Mahesh Malviya “Reduction of Negative and Positive Association Rule Mining and Maintain Superiority of Rule Using Modified Genetic Algorithm” International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December-2012.