An Optimization of Association Rule Mining using K-Map and Genetic Algorithm for Large Database

Ghanshyam Dhanore[1], Setu Kumar Chaturvedi[2]
Department of computer Science & Engineering[1, 2], TIT Bhopal (INDIA) [1, 2]

ABSTRACT
Rule mining is very efficient technique for find relation of correlated data. The correlation of data gives meaning full extraction process. For the mining of rule mining a variety of algorithm are used such as Apriori algorithm and tree based algorithm. Some algorithm is wonder performance but generate negative association rule and also suffered from multi-scan problem. In this paper we proposed a k-apriori-GA association rule mining based on genetic algorithm and K-map formula. In this method we used a k-map binary table for partition of data table as 0 and 1. The divided process reduces the scanning time of database. The proposed algorithm is a combination of k-partition and near distance of k-map candidate key. Support weight key is a vector value given by the transaction data set. The process of rule optimization we used genetic algorithm and for evaluate algorithm conducted the real world dataset The National Rural Employment Guarantee Act (NREGA) Department of Rural Development Government of India.

Keywords
Association rule mining, negative and positive rules, multi-pass, k-map, Genetic algorithm.

1. INTRODUCTION
The data mining also known as Knowledge-Discovery in Databases (KDD) and Data Mining is the method of involuntary searching large volumes of data for patterns using methods such as classification, association rule mining and clustering [1]. Data Mining is a difficult topic and has links with multiple core fields such as computer science and adds rate to wealthy seminal computational ability as of statistics, information or data retrieval, machine learning and pattern recognition. An association rule mining is a method which is used to perceive the unknown facts in huge dataset and draw interferences on how subsets of items influence the presence of other subsets. Association rule mining aims to find strong relation between attributes. All frequent generalized patterns are not extremely capable because a segment of the frequent patterns are superfluous in the association rule mining. That is why this algorithm produces some superfluous rule along with the interesting rule. Such drawback can be removed with the help of genetic algorithm. The description about association rule and genetic algorithm is described below:

1.1 Association Rules
Association rule mining [2] makes correlation among items that are grouped into transactions, inferring rules that describe the relationships among itemsets. The rules have a user specific support, confidence and length. An association rule is an implication of the form X->Y where X and Y are the item sets.

Support measures for metrics the fraction of transactions that surround both X and Y. Given a rule X->Y and N being the whole number of transactions then the support of an association rule is defined as:

Support=(X Union Y)/N

Confidence measures how frequently item in Y, appear in transactions that include X. Given the rule X ->Y its confidence is defined as follows:

Confidence= (X Union Y)/X

Itemsets with minimum support (minsup itemsets, though others that do not cross minimum support and minimum confidence values are known as small itemsets. Support is often used to eliminate the uninteresting rules. Confidence measures the reliability of the inference made by the rule. It also provides an estimation of the conditional probability of Y given X.

1.2 Genetic Algorithm
The genetic algorithm (GA) [3] keeps a population of n chromosomes (solutions) with allied fitness values. Parents are elected to mate, on the basis of their fitness, producing offspring using reproductive plans (mutation and crossover). Therefore, highly fit solutions are specified more opportunities to reproduce (selected for next generation). Subsequently, that offspring inherit characteristics from every parent. As parents mate and produce offspring, room have to be made for the new arrivals since the population is kept at a static size (population size). In this way it is hoped that over successive generations better solutions will succeed while the least fit solutions die out. The depiction schemes are Crossover rate, and fitness function and selection operator are the GA operators.

- **Fitness function:** The fitness of an individual in a genetic algorithm is the value of an objective function for its phenotype. For manipulating fitness, the chromosome has to be first decoded and the objective function has to be evaluated.
- **Selection:** Individuals are chosen from the current population since parents to be involved in recombination.
- **Crossover:** It is the course of taking two parent solutions and generating from them a two new offspring.

The application of association rule mining is on market basket data, whether prediction, multimedia data. In this paper optimization of large database association rule mining using k-map and genetic algorithm. The rest of paper is organized as follows. The organizations of the rest section are as follows: section II discusses related work of association rule mining. In section III description about the proposed algorithm and its block diagram and in section IV discuss experimental result followed by a conclusion and future work in Section V.

2. RELATED WORK
Various methods has been developed and proposed to optimize the database using association rules. In this paper some of the related method is described to extract the dataset using genetic algorithm:

1. **Nikhil Jain,Vishal Sharma,Mahesh Malviya:** Reduction of Negative and Positive Association Rule Mining and

Maintain Superiority of Rule Using Modified Genetic Algorithm [4] proposed the method in which find the near distance of rule set using Euclidean formula and produces two class advanced class and subordinate class. The authentication of class check through distance weight vector. Mainly distance weight vector sustain a threshold value of rule itemsets. Here the whole processes used the genetic algorithm for optimization of rule set. This proposed algorithm distance weight optimization of association rule mining with genetic algorithm compared with multi-objective association rule optimization using genetic algorithm.

2. **Peter P. Wakabi–Waiswa, Venansius Baryamureeba and Karunakaran Sarukesi**: Optimized Association Rule Mining with Genetic Algorithms [5] proposed an Association Rule Mining (ARM) technique to mine the large database which was introduced as a structured mechanism for finding the unknown facts in large datasets and drawing inferences on how a subset of items influences the presence of another subset.

3. **Sufal Das, Banani Saha:** Data Quality Mining using Genetic Algorithm [6] develop Multi-objective Genetic Algorithm (GA) based method utilizing linkage among feature selections or extraction and association rule. The major incentive for using GA in the discovery of high-level prediction rules is that they accomplish a global search and endure better with attribute interaction that the greedy rule induction algorithms frequently used in data mining.

4. **Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar:** Mining Frequent Item sets Using Genetic Algorithm [7] presents the Genetic algorithm (GA) is useful for large data sets to find out the frequent item sets. In this paper; they first load the model of records from the transaction database that well fitted into memory. The genetic learning starts as follows: an initial population is produced by consisting of randomly generated transactions. Each and every transaction can be symbolized by a string of bits.

The proposed genetic algorithm based method is used for finding frequent item sets continually transform the population by executing the following four stages:

i. **Fitness Evaluation:** The fitness or an objective function is deliberated for each individual.

ii. **Selection:** It is preferred from the current population as parents to be involved in recombination.

iii. **Recombination:** New individuals (called offspring) are produced from the parents by applying genetic operators such as crossover and mutation.

iv. **Replacement:** Some of the offspring are replaced with some individuals (usually with their parents). One cycle of transforming a population is called a generation. In each generation, a fraction of the population is replaced with offspring and its proportion to the entire population is called the generation gap (between 0 and 1).

5. Ashish Ghosh, Bhabesh Nath(2004):Multi-objective rule mining using genetic algorithms [8]. The author used metrics or measures like support, confidence, comprehensibility and interestingness for evaluating exciting rules as their dissimilar objectives for mining association rule task. Use of these measures they were implementing genetic algorithm to produce some relevant rules from any market basket analysis type database. On the basis of their experimentation the proposed algorithm has been found for optimization of large databases.

6. Bettahally, N. Keshavamurthy, Asad M. Khan, Durga Toshniwal: They proposed the Privacy preserving association rule mining over distributed databases using genetic algorithm [9]. In this author compares conventional frequent pattern mining algorithm i.e. Apriori algorithm with proposed genetic algorithm in local search. In the Apriori algorithm population is formed in only single recursion but in genetic algorithm population is formed in every new production. The algorithm also deals with various types of partitions such as flat, perpendicular, and random.

7. M. Ramesh Kumar and Dr. K. Iyakutti: Genetic algorithms for the prioritization of Association Rules [10] proposed a novel genetic algorithm based association rule mining algorithm is described in this paper. Prioritization of the rules has been discussed with the help of genetic algorithm. Fitness function is designed based on the two measures like all confidence and the collective strength of the rules, other than the classical support and the confidence of the rules generated. The algorithm has been tested for the four data sets such as Adult, Chess, Wine, and Zoo. They presented a fresh algorithm for the rule prioritizing, produced by the apriori algorithm through the application of genetic algorithm. The fitness function is deliberate based on the user's interesting measure and M is the threshold value of the interesting measure considered. The sample data sets have been taken from the UCI data repository for the testing of the algorithm. The proposed genetic algorithm based association rule mining algorithm for the prioritization of the rules. This approach significantly reduces the number of rules generated in the data sets. The fitness function is designed in such a way that to prioritize the rules based on the user preference.

8. Nidhi Sharma, Anju Singh: K-Partition Model for Mining Frequent Patterns in Large Databases [11] implemented the a K-Partition algorithm which impose the one database scan. The entire database is compacted in the form of Karnaugh map; contain very small size i.e. a fraction of the entire database. Then partition algorithm can be used to classify the frequent patterns using K-Map model. Thus this technique brings efficiency in terms of time taken by processor for mining frequent patterns.

3. PROPOSED ALGORITHM

This paper proposed a fresh algorithm for optimization of association rule mining, the proposed algorithm remove the problem of negative rule generation and also optimized the process of multi-pass of rules. Multi-pass of association rule mining is a great challenge for large dataset. In the generation of valid rules association existing algorithm or method generate a series of negative rules, which generated rule affected a performance of association rule mining.

In the process of rule generation various multi objective association rule mining algorithm are proposed but all these are not solve k-map problem of association rule mining. In this dissertation we proposed k-apriori-GA of association rule mining with genetic algorithm. In this algorithm we used k-map partition are used for logically partition of dataset. The database divided into section one is mapped data and another is scanned data. The mapped data logically assigned 1 and unmapped data logically assigned 0 for scanning process. The divided process reduces the scanning time of database.

The proposed algorithm is a combination of k-part ion and near distance of k-map candidate key. Support weight key is a vector value given by the transaction data set. The support value passes as a vector for finding a near distance between k-map candidates key. After finding a k-map candidate key the nearest distance divide into two classes, one class take a higher odder value and another class gain lower value for rule generation process. The process of selection of class also reduces the passes of data set. Once finding, a class of lower and higher of given support value, compare the value of distance wet vector. Now distance weight vector work as a fitness function for selection process of genetic algorithm. Now we present steps of process of algorithm step by step and finally draw a flow chart of complete process.

Steps of algorithm (K-Apriori-GA)

1. Select data set
2. Set value of support and confidence
3. Scan the transaction database to determine the support for each candidate item set
4. Logically divided dataset into two part 0 and 1
5. 0 assigned for mapped part and 1 assigned for unmapped part
6. Start scanning of transaction table
7. Count frequent items
8. Generate frequent item sets
9. Check the transaction set of data is null
10. Put the value of support as weight
11. Compute the distance with Euclidean distance formula
12. Generate distance vector value for selection process
13. Initialized a population set (t=1)
14. Compare the value of distance vector with population set
15. If value of support greater than vector value
16. Processed for encoded of data
17. Encoding format is binary
18. After encoding offspring are performed
19. Set the value of probability for mutation and the value of probability is 0.006
20. Set of rule is generated
21. Check k-map value of table
22. If rule is not k-map go to selection process
23. Else optimized rule is generated
24. Exit

Now we explain complete working process of an algorithm shown in as block diagram of proposed algorithm using k-map and genetic algorithm.

Figure 1 shows that the working process of proposed algorithm where dataset loaded at the initially and finally produced the reduced rules as an output. Block second describes the working of k-map where dataset is divided into subparts and proceed for the frequent item generation. And the whole process is completed according to the proposed algorithm block diagram (figure 1).

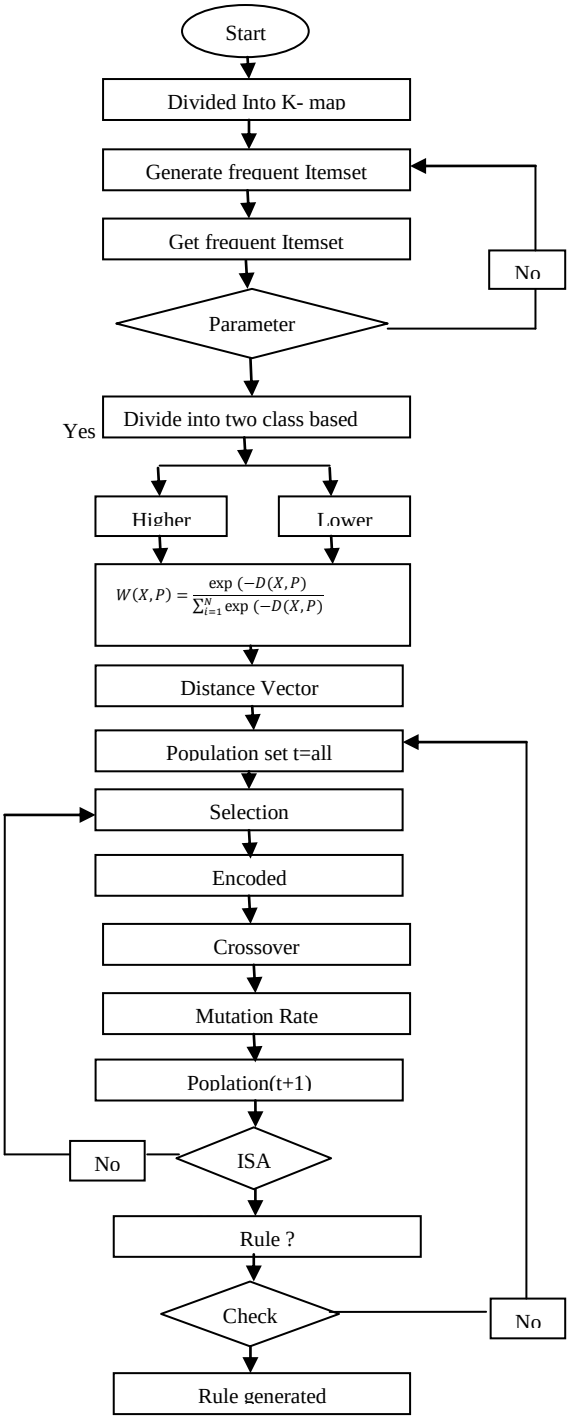

Figure1. Proposed block model

4. SIMULATION RESULTS

To investigate the effectiveness of the proposed method implemented in MATLAB 2009A [12] and for testing of result we used to evaluate algorithm conducted the real world dataset the national rural employment guarantee act (nrega) department of rural development government of India. The objective of the act is to enhance livelihood security in rural areas by providing at least 100 days of guaranteed wage employment in a financial year to every household whose adult members volunteer to do unskilled manual work proper maintenance of records is one of the critical success factors in the implementation of nrega. Information on critical inputs, processes, outputs and outcomes

have to be meticulously recorded in prescribed registers at the levels of district programme coordinator, programme officer, gram panchayat and other implementing agencies. The computer based management information system will also capture the same information electronically to evaluate algorithm used janpad panchayat march 2012 records from sehore, district (m.p.) details of the monthly squaring of accounts should be made publicly available on the internet at all levels of aggregation. The database contains data pertaining to name of sub engineers, name of gram panchayat, type of works, technical sanction, administrative sanction, cost of work, completion date, labour expenditure, material expenditure, total exp. of work, physical report of work in nrega has many recods but we randomly selected 731 transactions this transaction present 11 attributes and 731 instances. For theses experiments we have used 4 attributes and 20 instances and apply our proposed algorithm and pervious algorithm with varying support and confidence.

We have simulated apriori, k-apriori and k-aprioriGA respectively in matlab and figure 2 is the main GUI enviornment of the simulation windows, where on the top there are two input edittext box for get input minimum support and minimum confidence respectively at the very first stage.

Here figure 2 shows that the overall designing of the simulation environment and there are three pushbuttons use for three different methods, when we push the button then it count support and confidence for each generated rule, in the rule generation function, they used a variable rule_counter which increments when value of confidence is grater then support after each iteration, and when all rules finished then it shows the result of the current method to below of the current method was run.

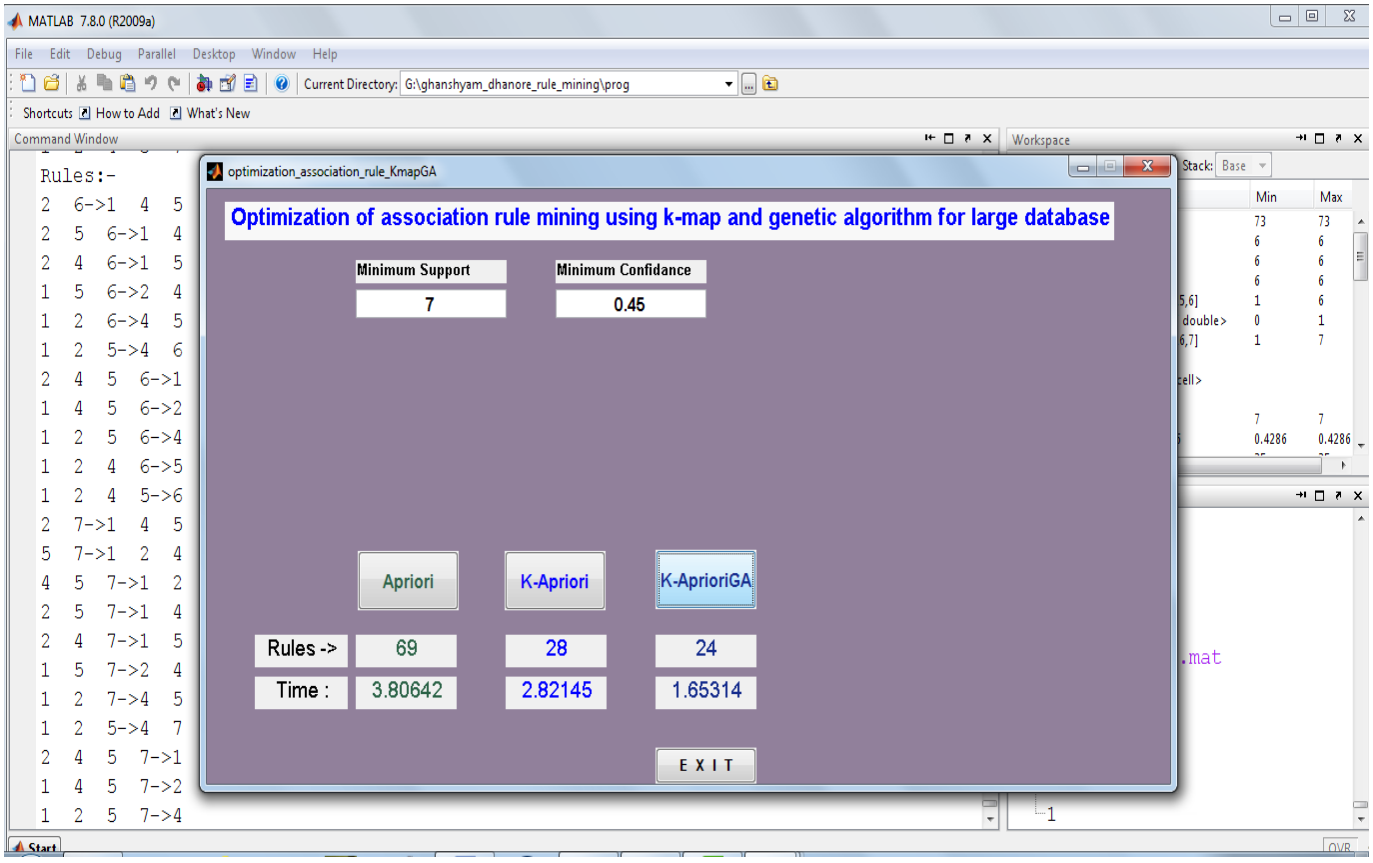

Figure 2: Main GUI Environment

Table1: Comparative result analysis of Apriori, K-Apriori and K-AprioriGA Methods

Minimum Support	Minimum confidence	Execution Time			Number of rules		
		Apriori	K-Apriori	K-AproriGA	Apriori	K-Apriori	K-AproriGA
2	0.15	2.90162	5.79219	**2.26628**	**57**	92	78
3	0.4	5.00763	2.56767	**1.37374**	75	**11**	**11**
4	0.22	4.80483	4.86663	**2.4836**	112	115	**107**
4	0.25	4.50843	4.80692	**2.21195**	110	**99**	**99**
5	0.35	3.24482	5.40405	**2.4836**	**50**	66	64
6	0.52	1.96561	2.10489	**0.93911**	24	22	**14**
7	0.45	3.80642	2.82145	**1.65314**	69	28	**24**
8	0.54	3.68162	4.32921	**1.84718**	46	43	**39**
9	0.65	2.58962	2.77667	**1.47464**	33	27	**23**
10	0.75	2.09041	2.43332	**1.28061**	13	17	**11**

Here table 1 show that the detailed analysis of the all the three algorithms with the different support and confidence values, where three methods generates their results on the two comparative parameters and best optimal result is indicated with the bold font. And in overall analysis of the comparative table the K-AprioriGA method gives the best result up to maximum time with other compared algorithms. Now we'll illustrate it with the help of bar graph.

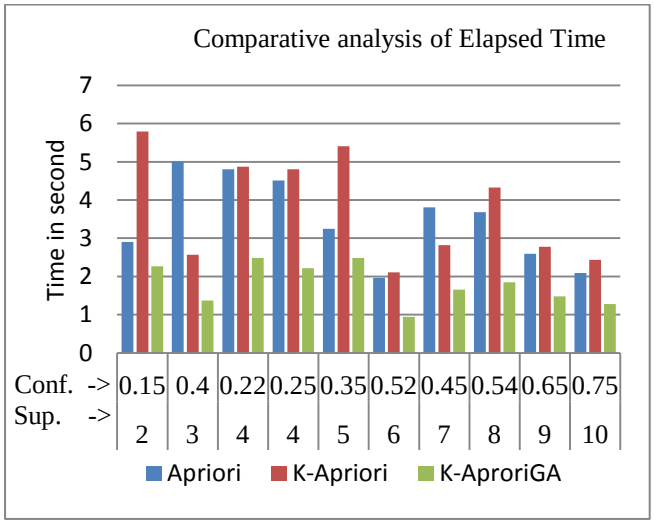

Figure 3: Comparission of elapsed time of three methods

Here figure 3 shows the comparative graph of elapsed time of all three methods where green bar indicate the K-AprioriGA method which show better performance or take less time to execution.

Now we'll illustrate the total rule generation of the all three methods with the help of bar graph in figure 4.

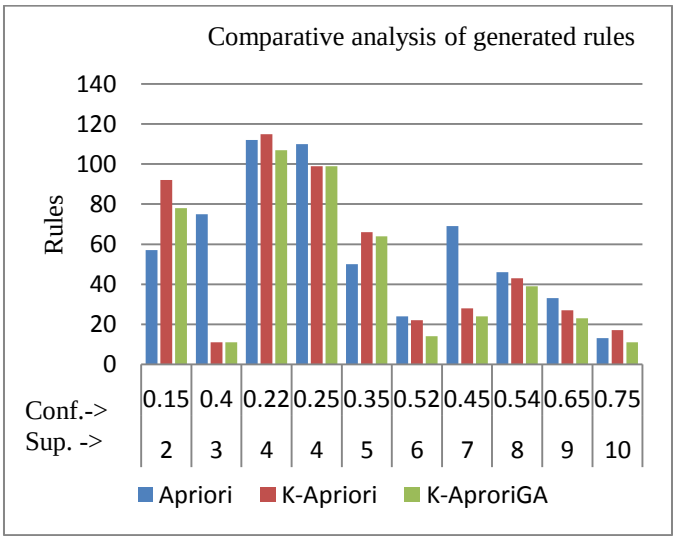

Figure 4: Comparission of generated rules of three methods

Here figure 4 shows that the comparative graph analysis of the all three methods in the respect of total rule generation in y-axis, and in x-axis indicate the varying values of the confidence and support. And the green bar indicting to the K-aprioriGA method, blue bar is for Apriori method and red bar is for K-Apriori method. And at finally we saw that the our proposed method K-aprioriGA shows the best result as compared to the other implemented algorithms.

5. CONCLUSION AND FUTURE WORK

This paper proposed a novel method for optimization of association rule mining. Our proposed algorithm is a combination of k-map and genetic algorithm. We have observed that when we modify the scan process of transaction generation of rule is fast. With more rules emerging it implies there should be a mechanism for managing their large numbers. The large generated rule is optimized with genetic algorithm. We theoretically proofed a relation between locally large and globally large patterns that is used for local pruning at each site to reduce the searched candidates. We derived a locally large threshold using a globally set minimum recall threshold. Local pruning achieves a reduction in the number of searched candidates and this reduction has a proportional impact on the reduction of exchanged messages.

6. REFERENCES

[1] Agrawal R., Imielinksi T. and Swami A. "Database mining: a performance perspective",(1993), IEEE Transactions on Knowledge and Data Engineering 5 (6), 914–925.

[2] R.Santhi and K.Vanitha: "An Effective Association Rule Mining In Large Database", International Journal of Computer Application and Engineering Technology Volume 1-Issue2, April, 2012 .pp72-76, ISSN: 2277-7962.

[3] Indira K and Kanmani S: "Performance Analysis of Genetic Algorithm for Mining Association Rules", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012 ISSN (Online): 1694-0814 www.IJCSI.org

[4] 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

[5] Peter P. Wakabi–Waiswa, Venansius Baryamureeba and Karunakaran Sarukesi "Optimized Association Rule Mining with Genetic Algorithms" in Seventh International Conference on Natural Computation, 2011.

[6] Sufal Das, Banani Saha: "Data Quality Mining using Genetic Algorithm", International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (2).

[7] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar: "Mining Frequent Itemsets Using Genetic Algorithm", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010 DOI: 10.5121.

[8] Jesmin Nahar a, Tasadduq Imama, Kevin S. Tickle a, Yi-Ping Phoebe Chen," Association rule mining to detect factors which contribute to heart disease in males and females ",Expert Systems with Applications, Elsevier , Vol. 40,pp.1086–1093,2013.

[9] Bettahally, N. Keshavamurthy, Asad M. Khan ,Durga Toshniwal, "Privacy preserving association rule mining over distributed databases using genetic algorithm", Neural Computing and Applications, Springer-Verlag, 2013.

[10] M. Ramesh Kumar and Dr. K. Iyakutti, "Genetic algorithms for the prioritization of Association Rules", IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications" AIT, 2011, pp. 35-38.

[11] Nidhi Sharma, Anju Singh: "K-Partition Model for Mining Frequent Patterns in Large Databases", International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397, Vol. 4 No. 09 Sep 2012.

[12] "MATLAB GUI". MATHWORKS. 2011-04-30. RETRIEVED 2013-08-14.