Mining Interesting Positive and Negative Association Rule Based on Improved Genetic Algorithm (MIPNAR_GA)

Nikky Suryawanshi Rai
PG Research Scholar (CSE),
RITS, Bhopal (M.P.) India

Susheel Jain
Asst Prof(CSE)
RITS, Bhopal (M.P.) India

Anurag Jain
HOD (CSE)
RITS, Bhopal (M.P.) India

***Abstract*—Association Rule mining is very efficient technique for finding strong relation between correlated data. The correlation of data gives meaning full extraction process. For the mining of positive and negative rules, a variety of algorithms are used such as Apriori algorithm and tree based algorithm. A number of algorithms are wonder performance but produce large number of negative association rule and also suffered from multi-scan problem. The idea of this paper is to eliminate these problems and reduce large number of negative rules. Hence we proposed an improved approach to mine interesting positive and negative rules based on genetic and MLMS algorithm. In this method we used a multi-level multiple support of data table as 0 and 1. The divided process reduces the scanning time of database. The proposed algorithm is a combination of MLMS and genetic algorithm. This paper proposed a new algorithm (MIPNAR_GA) for mining interesting positive and negative rule from frequent and infrequent pattern sets. The algorithm is accomplished in to three phases: a).Extract frequent and infrequent pattern sets by using apriori method b).Efficiently generate positive and negative rule. c).Prune redundant rule by applying interesting measures. The process of rule optimization is performed by genetic algorithm and for evaluation of algorithm conducted the real world dataset such as heart disease data and some standard data used from UCI machine learning repository.**

Keywords*—Association rule mining; negative rule and positive rules; frequent and infrequent pattern set; genetic algorithm

I. INTRODUCTION

Association rule mining is a method to identify the hidden facts in large instances database and draw interferences on how subsets of items influence the existence of other subsets. Association rule mining aims to discover strong or interesting relation between attributes. All generalized frequent pattern sets are not very efficient because a segment of the frequent pattern sets are redundant in the association rule mining. This is why, traditional mining algorithm produces some uninteresting rules or redundant rules along with the interesting rule. This problem can be overcome with the help of genetic algorithm. Most of the data mining approaches use the greedy algorithm in place of genetic algorithm. Genetic algorithm is produced by optimized result as compare to the greedy algorithm because it performs a comprehensive search and better attributes interaction [1]. In genetic algorithm population evolution is simulated. Genetic algorithm is an organic technique which uses gene as an element on which solutions (individuals) are manipulated. Generally association rule is used to finding positive relationship between the data set. Negative association rule is also vital in analysis of intelligent data. Negative association rule mining is adopted where a domain has too many factors and large number of infrequent pattern sets in transaction database. Negative association rule mining works in reverse manner and it define decision making capability, whether which one is important instead of checking all rules. However problem with the negative association rule is it uses huge space and can take more time to produce the rules as compare to the conventional mining association rule. In the generalized association rule database is scanned once and transaction is transformed into space reduced structure. The association rule mining problem can be decomposed in statistical and unconditional attributes in a database. The application of association rule mining is used to analyze various situation like market basket analysis, banks, whether prediction, pattern reorganization, multimedia data etc.

The process of optimization of interesting association rule mining used genetic algorithm. Genetic algorithm works in multiple levels of constraints for minimum support value and individual confidence value of frequent and infrequent patterns. The proposed method enhances the process of rule optimization for large datasets. The rest of paper is organized as follows. In Section II describes about related work of association rule mining. Section III describes about proposed method. Section IV describes about experimental result algorithm followed by a conclusion in Section V.

II. RELATED WORK

This section describes some related work to negative and positive association rule mining.

An Improved apriori algorithm is used minimum supporting degree and degree of confidence ,for extracting association rules .But it has suffered from “frequent pattern sets explodes ”and “rare item dilemma ” [2].

Improved multiple minimum support (MSapriori) based on notion of support difference and define how to deal with the problem caused by frequent pattern sets explodes ,but still suffer from rare item dilemma[3].

Primary stage of association rules, all algorithms based on single minimum support and those algorithms suffer from “rule missed” and “rule explosion” problem. An efficient

method to extract rare association rules .In this method the probability and introduces multiple minsupp value to discover rare association rules. One obstacle of this algorithm is that, it produces large number of uninteresting pattern sets [4].

PNAR_MDB on PS measures is introduced to discover PNAR in multi-databases. PNAR_MDB on PS extract interesting association rules by weighting the database (the weight of database must be determined) and used the correlation coefficient to remove the confliction of rules [5].

Reveal knowledge hidden in the massive database and proposed an approach for Evaluation of exam paper. This paper introduces a new direction, applies interesting rules mining to evolution of completive exam and finds out some useful knowledge. But this algorithm need repeatedly database scan and takes more time to perform I/O operation [6].

Some algorithm uses comparison support and comparison confidence (comsup, comconf) for extracting interesting relationship between pattern sets [7].

According to correlation and dual confidence measures association rules are classified in to positive and negative association rules ,but one drawback of dual confidence, is if less confidence would be a lot of rules even produce large number of contradict rules (¬C→¬D), if greater confidence may missed useful positive association rules[8].

Generalized Negative Association Rules (GNAR) is produced interesting negative rules ,this approach could speed up execution time efficiently through the domain taxonomy tree and extract interesting rules easily, advantage of taxonomy tree is to eliminate large number of useless transaction [9].

Another approach to solve key factors of interesting rules is PNAR algorithm, this algorithm efficiently define frequent pattern sets for interesting rules, NAR based on correlation coefficient and modified pruning strategy[10].

PNAR_IMLMS produces valid association rules based on correlation coefficient but one demerits of this algorithm, negative rules extract from uninteresting pattern sets which is useless [11]. Optimized association rule mining with genetic algorithm produces more reliable interesting rules compare to previous method.

Mining association rules using multiple support confidence values and several studies have been addressed the issue of mining association rules using Multiple Level Minimum Supports [12].

III. PROPOSED ALGORITHM

This paper proposed a novel algorithm for optimization of association rule mining, the proposed algorithm resolves the problem of negative rule generation and also optimized the process of rule generation. Interesting association rule mining is a great challenge for large dataset. In the generation of interesting rules association existing algorithm or method generate a series of negative rules, which generate rules which affect performance of association rule mining. In the process of rule generation various multi objective associations rule mining algorithm is proposed but all these are not solve.

This paper proposed an improved approach to mine association rule In this algorithm we used a MLMS for multi level minimum support for constraints validation. The scanning of database divided into multiple levels as frequent level and infrequent level of data according to MLMS. The frequent data logically assigned 1 and infrequent data logically assigned 0 for MLMS process. The divided process reduces the uninteresting item in given database.

The proposed algorithm is a combination of MLMS and genetic algorithm along this used level weight for the separation of frequent and infrequent item. The multiple support value passes for finding a near level between MLMS candidates key. After finding a MLMS candidate key the nearest level divide into two levels, one level take a higher odder value and another level gain infrequent minimum support value for rule generation process. The process of selection of level also reduces the passes of data set. After finding a level of lower and higher of given support value, compare the both values of level by vector function. Here level weight vector function work as a fitness function to define the selection process of genetic algorithm

Here we implemented the combinatorial method of MLMS and genetic algorithm for the mining of positive and negative item sets. The key idea is to generate frequent and infrequent item sets and with these item sets positive and negative association rules are generated. MLMS algorithm is used for the generation of rules [12], since the association rule mining seems to be better when the association rules are less, hence the minimization of these positive and negative rules can be done using genetic algorithm. The proposed technique can be described as follows:

1) Take an input dataset which contains number of attributes and instance values with single or multiple classes.

2) Initialize the data with length of the item sets k=2, 3, 4 and pass support and confidence (Para b).

3) Generate all the frequent and Infrequent item sets from MLMS algorithm for an item set of length k=2, 3, 4.

4) Generate positive association rules from frequent items sets and negative association rules from infrequent item sets.

5) Initialize all the general parameters involved in genetic algorithm.

6) Generate the child chromosomes of the positive and negative association rules and calculate the fitness value of each individual child chromosomes. Compare the individual fitness value of each child with the average fitness value and regenerate positive and negative association rules.

7) Crossover and mutate the remaining child chromosomes and reinitialize the fitness value and recalculate and regenerate final positive and negative rules.

A. Load Datasets

The association rules generated from the proposed algorithm needs datasets containing a number of transaction values. Here we use a number of datasets i.e. small and large dataset, a dataset with single and multiple classes. So the performance of the proposed methodology is tested for each datasets.

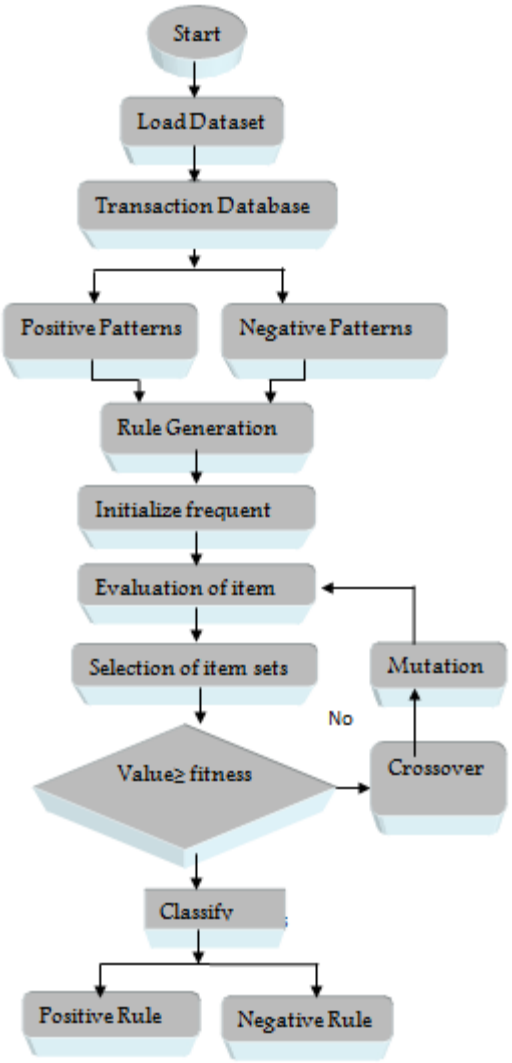

Fig. 1. Shows that proposed block model of algorithm.

B. Support and Confidence

Here the association rules can be generated on the basis of item set length, support and confidence. Suppose sup and cf are the support and confidence respectively. Let k be the length of the item set. For an item set A⊆I, the support is A.count / |TD|, where A.count is the number of transactions in TD that contain the itemset A. The support of a rule A⇒B is denoted as sup (A∪B), where A, B⊆I, and A∩B =Φ while the confidence of the rule A⇒B is defined as the proportion of s (A∪B) above s (A), i.e., cf (A⇒B) = s (A∪B) /s (A).

C. Generate Frequent and infrequent item sets

Here use MLMS algorithm for the generation of frequent and infrequent item sets. Form these frequent and infrequent item sets positive and negative association rules are generated.

A frequent itemset I: sup (I) ≥ minsupp
An infrequent itemset J: sup (J) ≤ minsupp

D. Correlation factor

For the generation of positive and negative association rules from these item sets, first of all correlation coefficient between the items sets is computed using:

$$corr_{AB} = \frac{cov(A,B)}{\sigma_A \sigma_B}$$

Where cov(A, B) represents the covariance of two variables and σ represents the standard deviation. Then compare the correlation coefficient with the correlation strength. Generate all the rules of the form

Positive association rules:

$$A \cap B = \phi$$
$$Supp\ (AU\ B) \geq minsupp$$
$$Supp\ (A\ U\ B)\ /\ supp\ (A) \geq minconf$$

Negative association rules:

$$A \cap B = \phi$$
$$Sup\ (A) >= minsupp,\ Sup\ (B) > minsupp,$$
$$and\ sup\ (A\ U\ \sim B) >= minsupp$$
$$Sup\ (A\ U\ \sim B)/sup\ (A) >= minconf$$

If the correlation coefficient is greater than or equal to α and if they meet the conditions VARCC(A,B,α,mc)=1 and VARCC(¬A,¬B,α,mc)=1. if the correlation coefficient is lower than or equal to -α and if they meet the conditions VARCC(A,¬B,α,mc)=1 and VARCC(¬A,B,α,mc)=1.

E. Initialization of Parameters

The genetic algorithm when applied should be initialized by certain parameters such as selection, crossover and mutation as well the number of iterations it will performed during working. There are various solutions that must be chosen randomly to form an initial population. The size of the population will depends on the problem

F. Fitness Function

The population selection for Genetic Algorithm is based on Fitness Function:

$$m(S) = \frac{Ai}{wi} + \frac{Bi}{L \times (1 - wi)}$$

Ai = {frequent item support}

Wi= {level of Wight value of MLMS}

Bi = {those value or Data infrequent}

The selection policy based on the foundation of individual fitness and concentration p(i) is the selection of individual whose fitness value is greater than one and m(s) is a value whose fitness is less than one but close to the value of 1.

The genetic operators find out the search capability and convergence of the algorithm.

G. Reproduction Operators

The child chromosomes that are not used in the sets will now be crossover and mutate so that the new fitness value is generated and again from parent, child chromosomes are generated. The process repeats until the rules generation finishes: Example:

1 0 1 0 0 1 0

↓

1 0 1 0 1 1 0

Mutation operator has been chosen to insure high levels of diversity in the population. We adopted PCA-mutation in (Munteanu 1999b), and shown that it has very good capabilities in maintaining higher levels of diversity in the population. We briefly summarize the PCA-mutation operator, as follows: The population **X** of the GA can be viewed as a set of *N* points in a *l*-dimensional space, where *N* is the size of the population and *l* is the length of the chromosome. It can be shown (Munteanu 1999b) that a GA converging has the effect of decreasing the number of Principal Components (PCs) as calculated with the Principal Components Analysis (PCA) method on data **X**.

a) Select a random point on the two parents.

b) Split parents at this crossover point.

c) Produce children's by exchanging trails.

d) Mutation typically in range (0.6, 0.9).

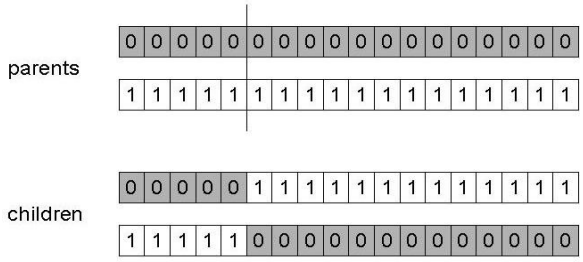

Fig. 2. Represent crossover of chromosomes

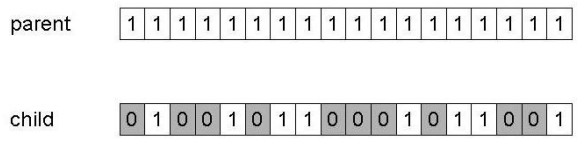

Fig. 3. Represent mutation of chromosomes.

IV. Simulation Result

This section shows the performance of MIPNAR_GA algorithm for mining both interesting positive and negative rules. Experiments are performed on a computer Intel Pentium dual core processor with 2.10 GHZ of CPU, running on a Windows 7 ,64-bit operating system and 4 GB of memory .All codes are implemented under the Java Compiler (JDK 1.6 and Weka 3.6.9) and Net Beans IDE version 6.9. Test the performance of proposed algorithm on 4 datasets from UCI machine learning website, which involve, Heart diseases, Breast Cancer, Wine and Iris. All information related to datasets are shown in Table 1.

TABLE I. Characteristics of Datasets

Dataset	No of Attributes	No of Instances	Classes
Heart Disease	14	303	2
Breast Cancer	10	286	2
Iris	14	178	3
Wine	5	150	3

Because MIPNAR_GA is designed to mine positive and negative rules from positive (frequent) and negative (infrequent) patterns with different input parameter (support, confidence, itemset length), it will be compared with the base algorithm PNAR_IMLMS for mining interesting positive and negative rules. The results are representing in table 2 to 7 where the number of interesting positive (A→B) and negative rules are represent as (A→¬B, ¬A→B, ¬A→¬B).

TABLE II. Show that given value of support (65%) confidence (55%) and item **length 2** algorithm PNAR_IMLMS generated total number of interesting Positive and Negative rules for UCI data set

Datasets	**PNAR_IMLMS**				
	A→B		**A→¬B**	**¬A→ B**	**¬ A→¬B**
Heart Disease	FIS	8	3	1	8
	inFIS	0	16	20	14
Breast Cancer	FIS	33	0	0	42
	inFIS	0	17	17	23
Iris	FIS	7	0	0	8
	inFIS	0	12	12	16
Wine	FIS	19	5	4	16
	inFIS	0	12	14	43
Total		**67**	**65**	**68**	**170**

TABLE III. Show value of support (75%) confidence (65%) and item **length 3** algorithm PNAR_IMLMS generated number of interesting Positive and Negative rules for UCI data set

Datasets	**PNAR_IMLMS**				
	A→B		**A→¬B**	**¬A→ B**	**¬ A→¬B**
Heart Disease	FIS	47	1	1	52
	inFIS	0	20	22	45
Breast Cancer	FIS	104	0	0	117
	inFIS	0	6	8	37
Iris	FIS	18	0	0	16
	inFIS	0	11	12	6
Wine	FIS	148	6	4	143
	inFIS	0	24	27	146
Total		**317**	**68**	**74**	**562**

TABLE IV. SHOW THAT GIVEN VALUE OF SUPPORT (55%) CONFIDENCE (45%) AND ITEM **LENGTH 4** ALGORITHM MIPNAR_GA GENERATED NUMBER OF INTERESTING POSITIVE AND NEGATIVE RULES FOR UCI DATA SET

Datasets	**PNAR_IMLMS**				
	A→B		**A→¬B**	**¬A→ B**	**¬ A→¬B**
Heart Disease	FIS	141	3	3	144
	inFIS	0	0	0	0
Breast Cancer	FIS	265	0	0	294
	inFIS	0	0	0	10
Iris	FIS	10	0	0	11
	inFIS	0	5	5	0
Wine	FIS	656	18	14	667
	inFIS	0	0	0	0
Total		**1072**	**26**	**22**	**1126**

TABLE V. SHOW THAT GIVEN VALUE OF SUPPORT (65%) CONFIDENCE (55%) AND ITEM LENGTH 2 ALGORITHM MIPNAR_GA GENERATED NUMBER OF INTERESTING POSITIVE AND NEGATIVE RULES FOR UCI DATA SET

Datasets	**MIPNAR_GA**				
	A→B		**A→¬B**	**¬A→ B**	**¬ A→¬B**
Heart Disease	FIS	3	0	0	2
	inFIS	0	7	9	10
Breast Cancer	FIS	12	0	0	16
	inFIS	0	6	6	8
Iris	FIS	2	0	0	3
	inFIS	0	5	5	8
Wine	FIS	10	1	0	9
	inFIS	0	8	8	16
Total		**27**	**26**	**28**	**72**

TABLE VI. SHOW THAT GIVEN VALUE OF SUPPORT (75%) CONFIDENCE (65%) AND ITEM LENGTH 3 ALGORITHM MIPNAR_GA GENERATED NUMBER OF INTERESTING POSITIVE AND NEGATIVE RULES FOR UCI DATA SET

Datasets	**MIPNAR_GA**				
	A→B		**A→¬B**	**¬A→ B**	**¬ A→¬B**
Heart Disease	FIS	31	0	0	35
	inFIS	0	10	12	18
Breast Cancer	FIS	75	0	0	80
	inFIS	0	1	2	9
Iris	FIS	5	0	0	8
	inFIS	0	4	5	1
Wine	FIS	130	2	2	112
	inFIS	0	16	16	97
Total		**241**	**33**	**37**	**360**

TABLE VII. SHOW THAT GIVEN VALUE OF SUPPORT (55%) CONFIDENCE (45%) AND ITEM LENGTH 4 ALGORITHM MIPNAR_GA GENERATED NUMBER OF INTERESTING POSITIVE AND NEGATIVE RULES FOR UCI DATA SET

Datasets	**MIPNAR_GA**				
	A→B		**A→¬B**	**¬A→ B**	**¬ A→¬B**
Heart Disease	FIS	97	1	0	98
	inFIS	0	0	0	0
Breast Cancer	FIS	203	0	0	215
	inFIS	0	0	0	2
Iris	FIS	3	0	0	4
	inFIS	0	1	1	0
Wine	FIS	598	7	5	602
	inFIS	0	0	0	0
Total		**898**	**9**	**6**	**921**

Table 2-7 shows the number of interesting positive and negative rules generated from useful positive and negative patterns with different input parameter. These rules are mined with two algorithms, the PNAR_IMLMS algorithm [12] and the MIPNAR_GA. For example, in Table 2 to 4 the number of interesting positive and negative rules mined by PNAR_IMLMS are 67 to 303 and 317 to 704 and 1072 to 1174, whereas in table 5 to 7 represent the total number of interesting positive and negative rules mined by MIPNAR_GA are 27 to 126 and 241 to 430 and 898 to 936 respectively .We can say that the algorithm MIPNAR_GA can successfully produce fewer rules than PNAR_IMLMS. In figure 3 to 5, P represent positive rule X→Y, N_1 represent A→¬B, N_2 represent ¬A→B, and N_3 represent ¬A→¬B.

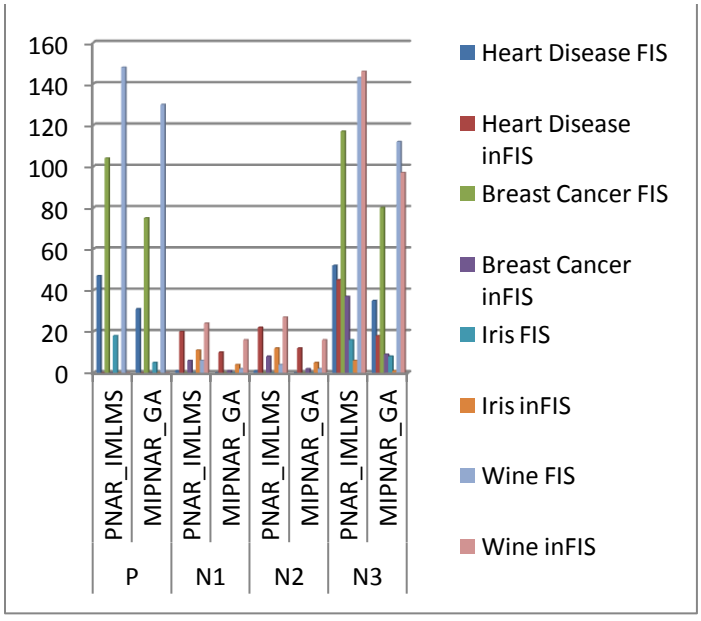

Fig. 4. Shows the comparative value of no of rules and reduces rules of two algorithms by optimization process from Tabel 3 to table 6.

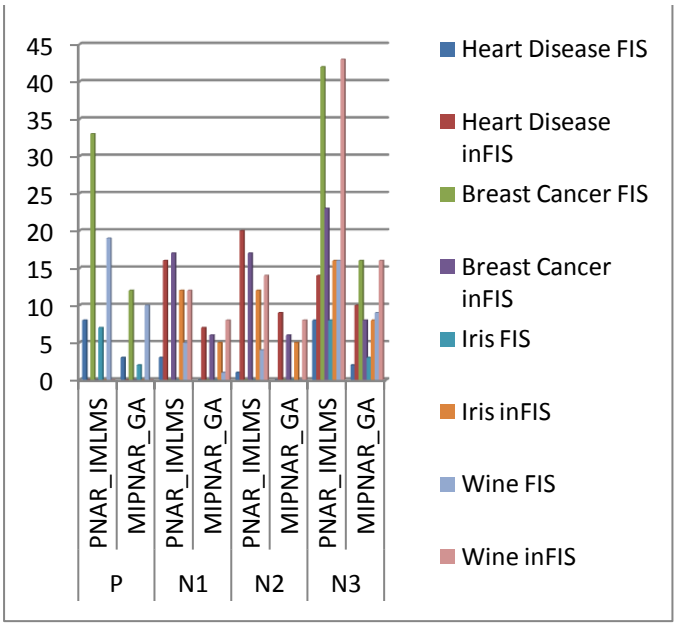

Fig. 5. Shows the comparative value of no of rules and reduces rules of two algorithms by optimization process from Tabel 2 to table 5.

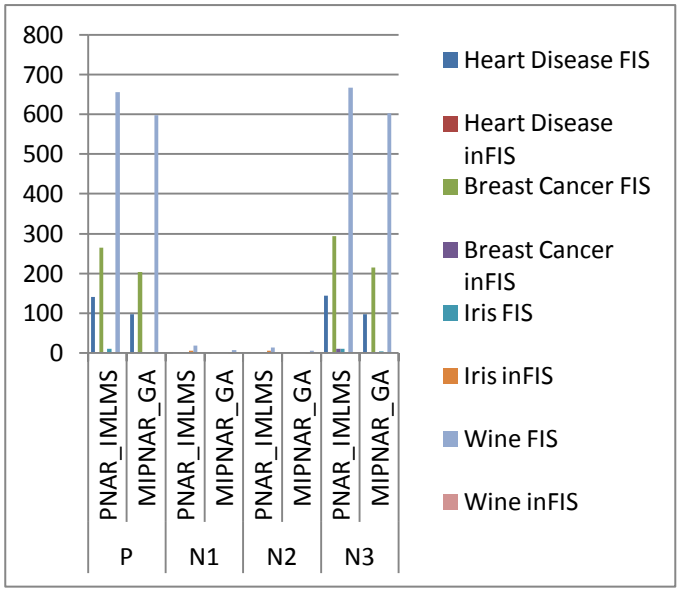

Fig. 6. Shows the comparative value of no of rules and reduces rules of two algorithms by optimization process from Tabel 4 to table 7

V. Conclusion And Future Work

This paper proposed a novel method for optimization of interesting positive and negative association rule. The defined algorithm is combination of MLMS and genetic algorithm. The observation is that when 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 pruning at each level to reduce the searched candidates. We derived a locally large threshold using a globally set minimum recall threshold. Pruning achieves a reduction in the number of searched candidates and this reduction has a proportional impact on the reduction of large number of negative rules. In future, some revision might take place to achieve two goals.

a) Various measures are added to this method for working with grid computing environment.

b) To improve the efficiency of the algorithm

Acknowledgment

I would like to thank Prof. Anurag Jain, Assistant Prof. Susheel Jain, for accepting me to work under his valuable guidance. He closely supervises the work over the past few months and advised many innovative ideas, helpful suggestion, valuable advice and support.

References

[1] By Pengfei Guo Xuezhi Wang Yingshi Han: "The Enhanced Genetic Algorithms for the Optimization Design" 978-1-4244-6498-2/10 © IEEE (2010).

[2] By WEI Yong-Qing, YANG Ren-hua, LIU Pei-yu: "An Improved Apriori Algorithm for Association Rules of Mining" 978-1-4244-3930-0/09/$25.00 © IEEE (2010).

[3] By R. Uday Kiran and P. Krishna Reddy:" An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules "978-1-4244-2765-9/09/$25.00 © IEEE (2009).

[4] By Sandeep Singh Rawat and Lakshmi Rajamani: "Probability Apriori based Approach to Mine Rare Association Rules".In 3rd Conference on Data Mining and Optimization (DMO), © IEEE (2011).

[5] By Shi-ju SHANG, Xiang-jun DONG, Jie LI, Yuan-yuan ZHAO: "Mining Positive and Negative Association Rules in Multi-database Based on Minimum Interestingness" 978-0-7695-3357-5/08 $25.00 © IEEE (2008).

[6] By XING Xue CHEN Yao WANG Yan-en:"Study on Mining Theories of Association Rules and Its Application" .International Conference on Innovative computing and communication Asia –Pacific Conference on Information Technology and Ocean Engineering 978-0-7695-3942-3/10 $26.00 IEEE (2010).

[7] By LI Tong-yan, LI Xing-ming: "New Criterion for Mining Strong Association Rules in Unbalanced Events" .Intelligent Information Hiding and Multimedia Signal Processing 978-0-7695-3278-3/08 $25.00 © IEEE (2008).

[8] By Xiufend Piao, Zhan long Wang, Gang Liu: "Research on mining positive and negative association rules based on dual confidence" Fifth International Conference on Internet Computing for Science and Engineering. 978-1-4244-9954-0/11 $31 © IEEE (2011).

[9] By Li-Min Tsai, Shu-Jing Lin and Don-Lin Yang: "Efficient Mining of Generalized Negative Association Rules" in International Conference on Granular Computing 978-0-7695-4161-7/10 $ 26 © IEEE (2010).

[10] By CH.Sandeep Kumar, K.Shrinivas, Peddi Kishor T.Bhaskar: "An Alternative Approach to Mine Association Rules" 978-1-4244-8679-3/11 $26.00 © IEEE (2011).

[11] By Dong, X., Niu, Z., Shi, X., Zhang, X., Zhu, D.: Mining both Positive and Negative Association Rules from Frequent and Infrequent Itemsets. ADMA, LNAI 4632, Springer-Verlag Berlin Heidelberg (2007)

[12] By Dong, X., Niu, Z., Zhu, D., Zheng, Z., Jia, Q:" Mining Interesting Infrequent and Frequent Itemsets Based on MLMS Model". The Fourth International Conference on advanced Data Mining and Applications, ADMA (2008).