An Efficient Methodological Study for Optimization of Negative Association Rule Mining

Piyush Vyas
Department of Information Technology
Shri Vaishnavsm Institute of Technology & Science
Indore,India

Jigyasu Dubey
Department of Information Technology
Shri Vaishnavsm Institute of Technology & Science
Indore,India

ABSTRACT
Association rule mining is interested area in present time for many research scholars. In this paper we show an efficient study of optimization of association rule mining. Lot of research done over positive association rule mining and now negative association rule mining is area of research. But we focus to study all kind of rules like positive and negative association rule as well as optimization of them with the help of genetic algorithm. Basically mining apply over Market basket, retail databases, and healthcare. Here we also study about basic Apriori algorithm for getting frequent item sets. Genetic algorithm also studied in deep in this research paper because of with the help of genetic algorithm we will found efficient association rules. In this research paper we also show some steps of methodology to find association rules through Genetic algorithm.

Key Words
Apriori Algorithm, Association rule mining, Genetic algorithm, Negative Association Rule.

1. INTRODUCTION
In recent years, Knowledge discovery in databases (KDD) has become a process of interest as the data in many databases have grown tremendously large. Pre-processing, data mining and post-processing are aspect of KDD. KDD provides opportunities to discovering useful information and important relevant patterns in large databases, that's why many researchers are primarily interested in it. There are many type of database so mining approaches may be classified as temporal database, relational database, multimedia database and transactional database. Data mining also classified in terms of mining classification rules, clustering rules, association rules and sequential patterns [1].

Association rule mining shows the pattern of data for transactional database. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc [2]. Consider a supermarket with a large collection of items. Typical business decisions that the management of the supermarket has to make include what to put on sale, how to design coupons, how to place merchandise on shelves in order to maximize the profit, etc. Analysis of past transaction data is a commonly used approach in order to improve the quality of such decisions. Until recently, however, only global data about the cumulative sales during some time period (a day, a week, a month, etc) was available on the computer [3].

Several organizations have collected massive amounts of such data. These data sets are usually stored on tertiary storage and are very slowly migrating to database systems. One of the main reasons for the limited success of database systems in this area is that current database systems do not provide necessary functionality for a user interested in taking advantage of this information [3].

Here we show in our study that we try to provide theoretically solution of association rule mining as well as to optimize them. Lot of association rules is garneted from Appriori algorithm so we show theory of optimize them through genetic algorithm.

2. LITERATURE SURVEY
Association rule mining among frequent items has been extensively studied in data mining research. However, in the recent years, there is an increasing demand of mining the infrequent items (such as rare but expensive items). A positive association rule is of the form: $X \Rightarrow Y$ and negative association rule is in the form $X \Rightarrow \neg Y$, $\neg X \Rightarrow Y$ and $\neg X \Rightarrow \neg Y$. Here X and Y is Item set of T transactional database.

Generally at starting age of data mining in 1993 first R. Agrawal, T. Imielinski and A. Swami invent about association rules mining between sets of items in large databases. They discuss about an efficient algorithm that generates all significant association rules between items in the database. they propose an algorithm to incorporates buffer management and novel estimation and pruning techniques in [4]. Then in 1994 again R. Agrawal and R. Srikant discuss about new fast mining algorithm for association rule mining. This time they concentrate over to generate new algorithm for find association rule fastly and efficiently. they presented two new algorithms, Apriori and AprioriTid, for discovering all significant association rules between items in a large database of transactions in [5]. after generation of association rule mining many researcher did lot of research in mining field and at starting of 2005 they came over positive and negative association rule mining and invent many algorithms . M.Jamali, F. Taghi yareh and M. Rahgozar talk about an encoding method to reduce the database size and after it by applying new algorithm on new layout they try to achieve significant efficiency in association rules discovery, they did research in 2005[6]. after many work over positive and negative association rule mining there are not a proper soluable platform for association mining than Chris Cornelis, Peng Yan, Xing Zhang and Guoqing Chen presented an Apriori-based algorithm that is able to find all valid positive and negative association rules in a support / confidence framework in [7]2006. Than in [10]2007, Sufal Das & Banani Saha tried to develop Multi-objective Genetic Algorithm (GA) based approach utilizing linkage between feature selection and association rule. They try

to inventing new area in mining for developing more efficient rule for transactional database. But at that time not many researcher concentrate over optimization of mining rules and again most of them work over positive and negative association rule mining. So, in 2008 [8] E.Ramaraj, N.Venkatesan worked over a new algorithm called BitArray-NegativePos that mines both positive and negative rules from the real time database. In 2008 [9] Honglei Zhu and Zhigang Xu work with a correlation coefficient measure and pruning strategies, their algorithm can find all valid association rules quickly and overcome some limitations of the previous mining methods. But at till coming 2009 there are lot of research done over positive and negative association rule mining then researcher again turn towards the optimization of association rule for example in 2009 [11] Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K work over optimization of association rule mining through ganatic algorithm. And in 2010 [12] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar and Partha Pratim Sarkar try to Mine Frequent Item sets Using Genetic Algorithm.

3. ASSOCIATION RULE MINING

Association rule mining falls under the descriptive category. Association rules aims in extracting important correlation among the data items in the databases. Association rule, basically extracts the patterns from the database based on the two measures such as minimum support and minimum confidence [13]. An association rule can be expressed as the form A $\Rightarrow$B, where A and B are sets of items, such that the presence of A in a transaction will imply the presence of B [17].

Two measures, support and confidence, are evaluated to determine whether a rule should be kept. The support of a rule is the fraction of the transactions that contain all the items in A and B. The confidence of a rule is the conditional probability of the occurrences of items in A and B over the occurrences of items in A. The support and the confidence of an interesting rule must be larger than or equal to a user-specified minimum support and a minimum confidence respectively [14] [22].As mentioned before association rule mining depends upon some basic terminologies like support count and confidence.

A. Support

Suppose we have nine transactions and five items. If an item set contain two or more item in a particular transaction than support count of an item is fraction of no. of that item comes in all transaction and total no. of transactions. Suppose support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item. The retailer will not pay much attention to such kind of items that are not bought so frequently, obviously a high support is desired for more interesting association rules.

$$\text{Supp}(A \Rightarrow B) = \text{supp}(A \cup B) = P(A \cup B) \quad (1)$$

In a single line we can say that "The support of an item-set is defined as the proportion of transactions in the data set which contain the item-set".

B. Confidence

The confidence of a rule is defined,

$$\text{Conf}(A \Rightarrow B) = \text{Supp}(A \cup B)/\text{Supp}(A) \quad (2)$$

Confidence can be interpreted as an estimate of the probability P(A |B), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS. In very simple way we say that, Confidence of an association rule is defined as the percentage or fraction of the number of transactions that contain A| B to the total number of records that contain A. Suppose confidence of the association rule A=>B is 80%, it means that 80% of the transactions that contain A also contain Y together.

C. Frequent and infrequent item set

A frequent item set is an item set that meets the user-specified minimum support. Accordingly, we define an infrequent item set as an item set that does not meet the user specified minimum support. Association analysis can be decomposed into the following two issues:

- Generate all item sets that have a support greater than or equal to the user specified minimum support are generated.
- Generate all the rules that have a confidence equal or greater than specified confidence.

D. Positive association rule

In this research work we also did research over positive association rule mining, so first we know about what is positive association rule? Some strong thing like if a customer buy milk he/she likely to by bread or if 80% of time customer buy milk and bread he/she likely to buy butter also in 20% of time. This rule is known as positive rule, due to example we conclude to things that milk and bread combination are really helpful to increasing productivity and butter is also a great option to make pair with bread for increasing productivity. Support calculation and confidence calculation for positive is shown earlier.

E. Negative Association rule

In association rule mining we evaluate negative association rule with the help of absence of item with regular item set. Suppose an item set contain A and B other contain B and C, so here in first item set C item is absent which show the negation of C. same as in second item set A is absent so, here shows the negation of A. due to this strategy we find out negative rules practically. Let understand theoretically all the possible negative rules. We categorized negative rule in three types, first is false positive, second is true negative and third is false negative.

- False Positive: $\neg A \Rightarrow B$, it means if A is not present but B.
- False Negative: $A \Rightarrow \neg B$, it means if A is present but not B.
- True Negative: $\neg A \Rightarrow \neg B$, it means if not A than also not B.

4. APRIORI ALGORITHM

In present scenario lot of association rule mining algorithms are available but Apriori algorithm is root or basic algorithm for generation of rules. Below we sited steps which fallow by algorithm to generate frequent item sets.

Step 1: Take a sample data set.
Step 2: Take user defines minimum support and confidence as an input.
Step 3: Apply Apriori algorithm;

- Calculate supports value of candidate.
- Calculate confidence value of candidate.
- First generate frequent item set through (Supp >= Min supp).
- After generating all the frequent item sets generate rules.

Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to counting the support of item sets and uses a candidate generation function which

exploits the downward closure property of support. In order to improve the algorithm efficiency, the Apriori property is introduced that all subsets of a frequent item set in DB are also frequent, and all supersets of an infrequent item set in DB are also infrequent [22]. Pseudo code of Apriori shown below,

Ck: Candidate item set of size k
Lk: frequent item set of size k
L1= {frequent items};
For (k= 1; Lk! =ǿ; k++) do begin
Ck+1= candidates generated from Lk;
For each transaction t in database do
Increment the count of all candidates in Ck+1that is
Contained in t
Lk+1= candidates in Ck+1with min_support
End
Return ∪k Lk;

Let we take an example and see how this algorithm generate frequent item sets step by step. Suppose below table 1 is transactional data set which shows no. of transaction as T.id and item sets. In it five item are present, I1, I2, I3, I4, I5. User specified minimum support is 2/9(means 2 support count is needed) (22%) and confidence is 70%. Table 2 and Table 3 shows the result of first step of Apriori algorithm, we get candidates item in from of C1(contain count of individual item means how many time they appear in whole trasactions) and frequent item in L1(contain item which satisfiy minimum support condition) with there support count.

TABLE I SAMPLE ITEM SET

Item set(L1)	**Support count**
I1	6
I2	7
I3	6
I4	2
I5	2

TABLE II CANDIDATE 1-ITEM

T id	**List of items**
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1,I2,I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3,I5
9	I1, I2, I3

TABLE III FREQUENT 1-ITEM

Item set (C1)	**Support count**
I1	6
I2	7
I3	6
I4	2
I5	2

TABLE IV CANDIDATE SET 2-ITEM

Item set (C2)
{I1,I2}
{I1,I3}
{ I1,I4}
{ I1,I5}
{ I2,I3}
{ I2,I4}
{ I2,I5}
{ I3,I5}
{ I3,I5}
{ I4,I5}

Here we complete first decomposition or itration to find out frequent item than in next itration algorithm go towards to find out 2-item set frequent pattern. In table 4, table 5, table 6 we clearly understand about 2 item set frequent pattern. In C2 every possible item combination got evaluated by scaning privious data. Table 5 also indicate support count of that items when they come toghether. In table 6 we got finally selected item set of 2 items after satisfing minimum support condition.

TABLE V CANDIDATE SET OF 2- ITEM WITH SUPPORT COUNT

Item set (C2)	Support count
{I1,I2}	4
{I1,I3}	4
{ I1,I4}	1
{ I1,I5}	2
{ I2,I3}	4
{ I2,I4}	2
{ I2,I5}	2
{ I3,I5}	0
{ I3,I5}	1
{ I4,I5}	0

TABLE VI FREQUENT SET OF 2- ITEM WITH SUPPORT COUNT

Item set (L2)	Support count
{I1,I2}	4
{I1,I3}	4
{ I1,I5}	2
{ I2,I3}	4
{ I2,I4}	2
{ I2,I5}	2

The generation of the set of candidate 3-itemsets, C3, involves use of the Apriori Property. In order to find C3, we compute L2JoinL2.

C3= L2 JoinL2 = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}.

Based on the Apriori property that all subsets of a frequent item set must also be frequent, we can determine that four later candidates cannot possibly be frequent. How? For example, lets take {I1, I2, I3}.The 2-item subsets of it are {I1, I2}, {I1, I3} & {I2, I3}. Since all 2-item subsets of {I1, I2, and I3} are members of L2, We will keep {I1, I2, and I3} in C3.Table 7, table 8, table 9 shows all final steps.

TABLE VII CANDIDATE SET OF 3-ITEM

Item set(C3)
{I1,I2,I3}
{I1,I2,I5}

TABLE VIII CANDIDATE SET OF 3-ITEM WITH SUPPORT COUNT

Item set(C3)	Support Count
{I1,I2,I3}	2
{I1,I2,I5}	2

TABLE IX FREQUENT SET OF 3-ITEM WITH SUPPORT COUNT

Item set(L3)	Support Count
{I1,I2,I3}	2
{I1,I2,I5}	2

The algorithm uses L3 JoinL3to generate a candidate set of 4-itemsets, C4. Although the join results in {{I1, I2, I3, I5}}, this item set is pruned since its subset {{I2, I3, I5}} is not frequent. That's why algorithm run till 3 iterations and did 3 level of decomposition.

5. OPTIMIZE ASSOCIATION RULES

A. Genetic Algorithm

Genetic Algorithm (GA) [10] was developed by Holland in 1970. This incorporates Darwinian evolutionary theory with sexual reproduction. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution. GA has been successfully applied in many search, optimization, and machine learning problems. GA process in an iteration manner by generating new populations of strings from old ones. Every string is the encoded binary, real etc., version of a candidate solution. An evaluation function associates a fitness measure to every string indicating its fitness for the problem. Standard GA apply genetic operators such selection, crossover and mutation on an initially random population in order to compute a whole generation of new strings [10].

- A chromosome is a set of parameters which define a proposed solution to the problem that the genetic algorithm is trying to solve [10].
- It describes the ability to both survive and reproduce, and is equal to the average contribution to the gene pool of the next generation that is made by an average individual of the specified genotype or phenotype [10],[18],[19].
- Randomly selected chromosomes are further crossing over according to their fitness value means which chromosome has maximum fitness value participating in crossover for create new strong generation of solutions [16].
- After crossing over if newly generated populations evaluate same fitness value continuously than GA call mutation for new generation. We use flipping mutation in it simply 0 replace by 1 and 1 by 0 [21].
- When maximum no of generation take place or no. of mutation crossover completed than GA stops.

After Appling these all process over association rules we got less no. of optimal solutions but it takes time and extra processing so, to reduce this problem we originate our modified genetic algorithm which we discuss in further section [21].

B. Approach to find Association rules

In this section we describe our experimental algorithm though which we get association rules. Here we discuss about

our fitness function which we build for getting all positive and negative rules. In below showing approach rules_sup indicates original support values of all rules which evaluated after applying Apriori algorithm and min_sup indicates user specified support values which we provide beginning of Genetic algorithm. Rules_conf indicates original confidence values of all rules which evaluated after applying Apriori algorithm and min_conf indicates user specified confidence values which we provide beginning of Genetic algorithm.

Step 1: Generate initial population randomly.

Step 2: Compute fitness of each individual

```
If (rules_sup >= min_sup) && (rules_conf >=
 min_conf)
        Fitness = rules_sup*rules_conf;
End;
```

Step 3: Perform no. of Crossover (mating).

Step 4: Perform no. of Mutation for creating new population.

Step 5: If no. of Crossover and no. of Mutation completed than finish Genetic algorithm.

6. CONCLUSION & FUTURE WORK

This research paper flow towards the study of optimal solution of association rule once again we understand about the basic approach of this research paper. First we study about Apriori algorithm over sample data to get association rules than to get optimal solution and important rule we studied genetic algorithm. In future we will use other large data base of different fields and try to do other modifications in basic studied approach to reduce other complexities like performance majors. We also will try to make robust discussed fitness function and apply other optimization techniques.

7. REFERENCES

[1] Qiankun Zhao: Association Rule Mining: A Survey, Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003.

[2] Sotiris Kotsiantis, Dimitris Kanellopoulos, Association rules mining: A recent overview, International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[3] Rakesh Agrawal Tomasz Imielinski_ Arun Swami, Mining association rules between sets of Items in large databases, ACM SIGMOD Conference Washington DC, USA, May 1993.

[4] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the Association for Computing Machinery—Special Interest Group on Management of Data, ACM-SIGMOD, May 1993, pp. 207–216.

[5] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Databases, VLDB, September 1994, pp. 487–499.

[6] M.Jamali, F. Taghi yareh, M. Rahgozar, Fast Algorithm for Generating Association Rules with Encoding Databases Layout, in: World Academy of Science, Engineering and Technology 4 2005.

[7] Chris Cornelis, Peng Yan, Xing Zhang, Guoqing Chen, Mining Positive and Negative Association Rules from Large Databases, in: IEEE 2006.

[8] E.Ramaraj, N.Venkatesan, Positive and Negative Association Rule Analysis in Health Care Database, in: IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.10, October 2008.

[9] Honglei Zhu, Zhigang Xu, An Effective Algorithm for Mining Positive and Negative Association Rules, in: International Conference on Computer Science and Software Engineering, 2008.

[10] Sufal Das & Banani Saha, Data Quality Mining using Genetic Algorithm, in: International Journal of Computer Science and Security, (IJCSS) Volume (3) 2007: Issue (2), pp. 105-112.

[11] Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K. Optimized association rule mining using genetic algorithm, in : Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, 2009, pp-01-04.

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

[13] Maragatham G, Lakshmi M, A recent review on association rule mining, Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 6 Dec 2011-Jan 2012.

[14] Yeong-Chyi Lee a, Tzung-Pei Hong b,_, Wen-Yang Lin c, Mining association rules with multiple minimum supports using maximum constraints, Preprint submitted to Elsevier Science 22 November 2004.

[15] B.Kavitha Rani1, K.Srinivas2, B.Ramasubba Reddy3, Dr.A.Govardhan4, Mining Negative Association Rules, International Journal of Engineering and Technology Vol.3 (2), 2011,pp 100-105.

[16] Dr (Mrs).Sujni Paul, An optimized distributed association rule mining algorithm in parallel and distributed data mining with xml data for improved response time, International Journal of Computer Science and Information Technology, Volume 2, Number 2, April 2010. pp 88-101.

[17] Yun Sing Koh1 Russel Pears2, Rare association rule mining via transaction clustering, Conferences in research and practice in Information Technology (CRPIT), Vol.87, 2008.

[18] Virendra Kumar Shrivastava Dr. Parveen Kumar Dr. K. R. Pardasani, Extraction of interesting association rules using GA optimization,global Journal of Computer Science and Technology Vol. 10 Issue 5 Ver. 1.0 July 2010, pp 81-84.

[19] Asst. Prof. Nirupama Tiwari Anubha Sharma1, A Survey of association rule mining using genetic algorithm, National Conference on Security Issues in Network Technologies (NCSI-2012) August 11-12,2012

[20] Badri Patel, Vijay K Chaudhari, Rajneesh K Karan, YK Rana, Optimization of association rule mining Apriori algorithm using ACO, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-1, March 2011, 24-26.

[21] S.N. Sivnandam, S.N. Deepa, Principles of Soft computing, (Wiley india pvt. Ltd, New delhi, 2011), pp 385-464.

[22] Badri Patel, Vijay K Chaudhari, Rajneesh K Karan, YK Rana: Optimization of Association Rule Mining Apriori Algorithm Using ACO, International Journal of Soft Computing and Engineering (IJSCE),ISSN: 2231-2307, Volume-1, Issue-1, March 2011, pp. 24-26.