International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 05, May-2015

A modified Semi-Apriori Algorithm to mine Frequent and Rare itemsets using Multiple Minimum Support

Gandhi Priyank Sanjay
LJIET,Gujarat Technological University
Ahmedabad,Gujarat,India

Ms. Jasmin Jha
LJIET,Gujarat Technological University
Ahmedabad,Gujarat,India

Abstract:- **The extraction of hidden information which can be predicted from large databases is known to as Data Mining. It's a new and powerful technology which help companies to focus on the most important data as well as information in their data warehouses. It basically involves the process of discovering hidden values in your data ware house. The tools related to Data mining are tend to predict future trends and behaviors which allows business and industries to make knowledge driven decisions. Today to understand various patterns, most of the companies collect refined mass quantity of data. The data mining technique can be implemented on existing hardware and software platforms enhances the value of current existing information resources. The present mining algorithms cannot perform efficiently due to high and repeatedly database scan. The association rules containing rare items are rare association rules. The less frequent items are known to be rare items. Mostly real world datasets are non uniform in nature and contains both frequent and rare occurring entities. It's more difficult to detect and generalize rare cases as they contain fewer data. It's important to understand the importance of rare knowledge patterns pertaining to rare events. Various research works are going on to investigate the more improved approaches for extracting rare knowledge patterns like rare association rules and rare class identification.**

Keywords:-Data Mining, knowledge patterns, database, data, rare association rule

Overview:- The extraction of hidden information which can be predicted from large databases is known to as Data Mining[8]. It's a new and powerful technology which help companies to focus on the most important data as well as information in their data warehouses. It basically involves the process of discovering hidden values in your data ware house. The tools related to Data mining are tend to predict future trends and behaviours which allows business and industries to make knowledge driven decisions. Today large quantity data is collected to understand various patterns, by the companies. To enhance the value of existing information resources, data mining techniques can be used on current hardware and software platforms. As they are brought online, the data mining techniques can be easily integrated with new systems and products. When this is implemented on high performance client-server or parallel processing computers, the data mining tools analyses large databases to answer various questions such as, "Which clients would mostly respond to the next promotional offers, and why ? "

From the name itself its known to have similarities in searching valuable business information from various large databases. For example, in understanding how many patients discharged from the hospital would be getting infected and are re-admitted how many cars which fueled up at the same gas station have got their owners stuck on the roadside. Both these processes require to go through immense material and find out what exactly had happened and which values resides. If the database is of appropriate quality and size, the technology for data mining can for sure generate new business opportunities.

- *Prediction of automated trends and behaviours*. It's the process for prediction of information from large databases . Questions which traditionally required to be analyzed can be now directly answered from the data. Example of this predictive problem is targeted marketing. The past data of promotional mailings are thoroughly checked for targeting the loyal customers or those who have expressed their interest on the same to get maximum return on investment for future mailings.

- *Discovery of automated previously unknown patterns.* It's the data mining tools which scans through databases and identify the previously hidden patterns in a single step. The example related to this can be the analysis related to retail sales which identifies which unrelated or rare products are often purchased together. The example related to this can be how many person who bought bed also bought side tables along with it.

The data mining techniques can benefit the business owners the automation and up gradation on existing and new hardware or software platforms. The massive databases can be analyzed in minutes if the data mining tools are implemented on parallel processing computers and high performance client-server. This results in faster processing. It means that the user himself can experiment automatically for the more models in order to understand complex data.[9] Its only the high speed which makes this much efficient. Larger databases gives improved predictions.

Discovery of relationships among itemsets in a transactional database is the main goal of association rule mining. Association rule mining was introduced by Agrawal et al. (1993)[9]. The aim of rare association rule mining is to extract interesting frequent patterns, correlations, casual structures or associations among itemsets in data repositories or transaction databases. The inherent properties of the data does not depend only on the relationships. The co-occurrence of the items in the database are dependent on relationship. Association rules are known to be what associations are there between items. From frequent items association rules are derived, the representation is done in the form of A→B where AB is a frequent itemset . The Strong association rules can be identified with those that meeting the minimum confidence c threshold (the percentage of transactions which contains A also contains B). All frequent itemsets are found, where an itemset is said to be frequent if it appears with minimum frequency s, called minimum support, in the traditional association rule mining process.

Mining infrequent itemset or rare association rule mining is much less explored area in association. We pruned out the items which occur rarely or are in very few transactions. The limitation of normal association rule mining approaches, i.e. Apriori, relies on being a meaningful minimum support level and it's reasonable (sufficiently strong) to reduce the number of frequent itemsets generated to a certain manageable level[9]. The relatively infrequent associations are likely to be of great interest as they relate to rare but crucial cases in some data mining applications. Examples which are based on rare association include predicting telecommunication equipment failure, identifying relatively rare diseases, and finding associations between infrequently purchased items in supermarket. Rare itemsets require special attention as they are more difficult to find by using traditional data mining techniques.

Problem Situation:- The greatest difficulty is the setting of support threshold in applying association rules mining. Assumption that that all items in the database are of the same kind and have similar frequencies in the database is not valid in reality . Some items are appearing very frequent in the database. Some other items hardly ever appear. Moreover only the frequent itemsets are alone not interesting. High support and high confidence thresholds generate frequent associations. Also low support and high confidence association need to be generated, which are rare. These Rare association may express information of high interest to experts and business people.

There are mainly two sub problems associated with association rules problem. The first is the discovery of the itemsets and their occurrences that go beyond a predefined minimum threshold. If this condition is satisfied then it will be known as large or frequent items. Second problem is using those large frequent items for generating large frequent items to generate association rules with the constraints of minimum confidence. Here we tend to consider the problem of finding rare as well as frequent itemsets because its necessary and computationally expensive.

Proposed System:- The proposed algorithm is an improved semi apriori approach to extract frequent and rare itemsets discovering rare association rules. Here, the notion of "support difference" is calculated to find minimum support for each item. This algorithm dynamically assigns appropriate minimum support to each item. By doing so the frequent itemsets involving rare items can be extracted in a more efficient manner as compared to the existing approaches. Most important, the proposed approach makes sure that the difference between the support of an item and the corresponding minimum support remains constant for all frequent as well as rare items. This reaults in efficiently reducing the explosion of frequent itemsets involving frequent items that also without affecting the extraction of frequent itemsets involving rare items

Problem Background:- Still, mining the frequent as well as rare itemsets together is a major challenge in data mining. The frequent itemsets generation produce extremely large numbers of generated itemsets that makes the algorithm inefficient. The only reason for this is that most approach uses traditional iterative strategy to discover itemsets which require very large process. Also present mining algorithms are not efficient as they perform high and repeated database scan. The single minimum support (minsup) based approaches like Apriori is used for extracting infrequent itemsets. It suffers from "rare item problem" dilemma. When the minsup value is high,rare itemsets re missed and when kept low, we get many frequent itemsets. If we set high percentage value by fixing the minsup, rare itemsets are missed as the minsup of rare items becomes close to their support. If the low percentage value of the minsup is fixed, there's flood of frequent itemsets.

Related Work:- The notion of finding rare association rules is like finding precious gems in an open field. It's a daunting task but, if successful, it is very rewarding. ARM systems, such as Apriori, generally employ an exhaustive search algorithm. These algorithms are capable of finding rare association rules. if the minimum level of support is set low enough to find rare rules they become intractable. For finding rare associations, such algorithms are therefore inadequate and also suffer from the rare item problem. Research to solve this problem has become more prevalent in recent times. To discover relationships among sets of items in a transactional database that occur infrequently is the major goal of rare association rule mining.

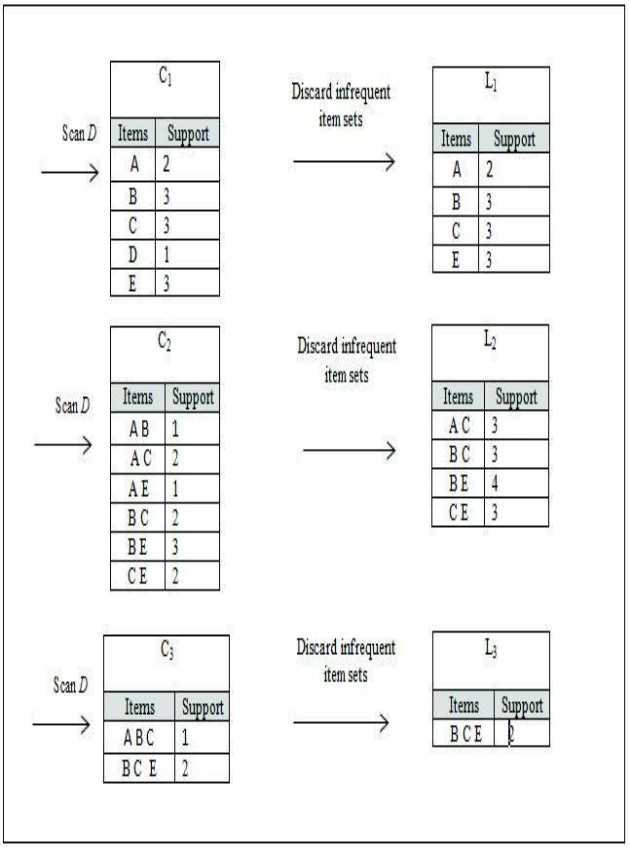
- *Apriori Algorithm*
- Apriori is a one of the first, famous, well known, and scalable algorithm for mining frequent itemsets and association rule.
- The algorithm works on the process of searching large itemsets during its initial database pass and uses its result for the other large datasets during subsequent passes.

- Rules which have low support level than the minimum support threshold are known as infrequent itemsets and those items which have support more than or equal to minimum support, will be known as large frequent itemsets.
- The problem of Apriori algorithm is, it requires multiple database scan, and additionally generates many candidates.
- Below example illustrates the overall idea of Apriori algorithm.

TID	Items Used
100	A,C,D
200	B,C,E
300	A,B,C,E
400	B,E

Table 1:- Sample contents of database transactions

Figure 1



Apriori frequent item mining steps[1]

- *Improved Multiple Support Apriori Algorithm (IMSApriori).*

- Its an improved approach in which minsup is fixed for each item based on the notion of "support difference".
- This approach improves performance over single minsup based approaches

- The assignment of appropriate minsup values for frequent as well as rare items is based on their item supports and reduces both "rule missing" and "rule explosion" problems.

- Support difference (SD):- Calculates minimum support(minsup). It extracts the frequent itemsets which involves rare items also and limits the explosion of frequent itemsets which involve frequent items

- Proposed Methodology :- Improved Multiple Support Apriori Algorithm (IMSApriori).

1. Generate Large Itemset
2. Calculate Minimum Item SupportItem Support.

- *Mining Rare Association rules in a Distributed Environment*

- It utilizes the idea of using statistic percentile which helps to produce multiple minimum supports to mine rare association rules in distributive environment.

- According to different characteristics of item sets support, different minimum supports for each level of item sets.

Table 2

A	B	C	E
2	3	3	3

- More rare association rules with an optimized communication cost can be derived here.
- Proposed methodology :

1. Apriori_MSG-P Algorithm.(Different Minimum supports are created by a user specific percentile value))
2. AprioriMSD(Find rare item sets)
3. CMS(Calculate Minimum Support)

- *Semi Apriori Algorithm*

- The Semi-Apriori algorithm for mining frequent items is divided into three stages.
- Instead of Table 2,transaction is represents in binary manner

Table 3: Transaction items in a binary representation

TID	A	B	C	D	E
100	1	0	1	1	0
200	0	1	1	0	1
300	1	1	1	0	1
400	0	1	0	0	1

- The first stage starts by finding the 1-itemsets L1 and pruning all items that have support less than the given minimum support threshold. This step is similar to the step used in Apriori and FP-Growth algorithms
- The first stage output can be shown in table (3)

Table (3) First frequent1-itemset[1]

- In the second stage, the algorithm figure (3) applies self-joint of L1 using L1⋈L1, also using the support measure. The items which are below the minimum support will be pruned. The second stage output can be shown in table (4)

AC	BC	BE	CE
2	2	3	2

Table 4;- Second frequent 2-itemset[1]

- In the third stage, the frequent itemsets of size > 2 are generated.

TID	Items Combinations
100	A, C, D, AC, AD, CD, ACD
200	B, C, E , BC, BE, CE, BCE
300	A,B,C,E,AB, AC, AE, BC, BE, CE, ABCE
400	B, E, BE

Tables 5:- actual combinations[1]

- Table (5) below reveals all actual combinations that occur within the transactions given in table (2). For example, in transaction T100, only three items are non-zero which are ACD. Thus, the combination of T100 will be A alone, C alone, D alone, AC, AD, CD, and ACD

Comparison among various related work

Sr No.	Parameters	Approach	Pros	Cons
1	On the Use of Ant Programming for Mining Rare Association Rules	The GBAP-RARM Algorithm	High complex problem can be addressed disregard to no. of attributes or instances , uniform processing time	Huge no. of rules extracted. Difficult to interpret and manage.
2	A Fast Algorithm for Mining Rare Itemsets	Rarity Algorithm	Faster	High Memory Usage
3	Mining Rare Association Rules in a Distributed Environment using Multiple Minimum Supports	1. Apriori_MSG-P 2. AprioriMSD 3. CMS	AprioriMSD is an optimal distributed data mining algo. Recognizes Skewed datasets and handles such datasets appropriately. Optimized comn. cost	Higher size of item sets. High memory usage.
4	An improved Multiple Minimum Support Based Approach to Mine Rare Association Rules	An improved Multiple Minimum Support Based Approach to Mine Rare Association Rules	Improves performance,finds frequent itemsets involving rare items.	Iterative approach makes it time consuming.
5	An Efficient Approach to Mine Rare Association Rules Using Maximum Items Support Constraints	MCCFP-Growth	Single scan on transactional dataset	MIS(Minimum Item Support) has to be defined by the user.
6	A Semi-Apriori algorithm for discovering the frequent Itemsets	Semi Apriori Algorithm	Avoids repeateed scans,avoids generation of large number of candidate sets,minimize execution time.	Cannot have multiple minimum support for every item

Table 6: Comparison among various related work

Proposed System :-

To overcome the problems associated with association rules problem, we tend to find rare as well as frequent itemsets with the semi apriori approach. First of all what is to be done is to have the data represented in the binary form. This uses binary mapping and reusing the previous data. The notion of support difference SD to specify the minimum supports to items is used here. acceptable deviation of an item from its frequency (or support) is refered by Support difference (SD). By this an itemset involving that item can be considered as a frequent itemset. For each item 'i', calculation of minsup known as minimum item support (MIS(i)) is as follows:-

$$MIS(i)=LS\ ,$$

where, LS refer to the least support which is specified by user.

Using SD, MIS for the items range from $(-\infty,+\infty)$.We use concept of least support (LS) to prevent MIS values of the items in reaching 0 or lower. the lowest minimum support an item or itemset should satisfy to become a frequent itemset is refered Least support. [0%, 100%] value is the range that LS takes.

The presented proposed semi apriori approach generates frequent and rare itemsets by taking whole transaction and increasing the frequency of each item appeared in the transaction by one. However, the contrary to single minsimum support approach that follow "downward closure property" (all the subsets of a frequent itemset are frequent), while multiple minimum support approaches follow "sorted closure property". Item cannot be discarded as any addition of items to it can be frequent,if an itemset is not frequent at (k-1) itemset,

Flowchart:-

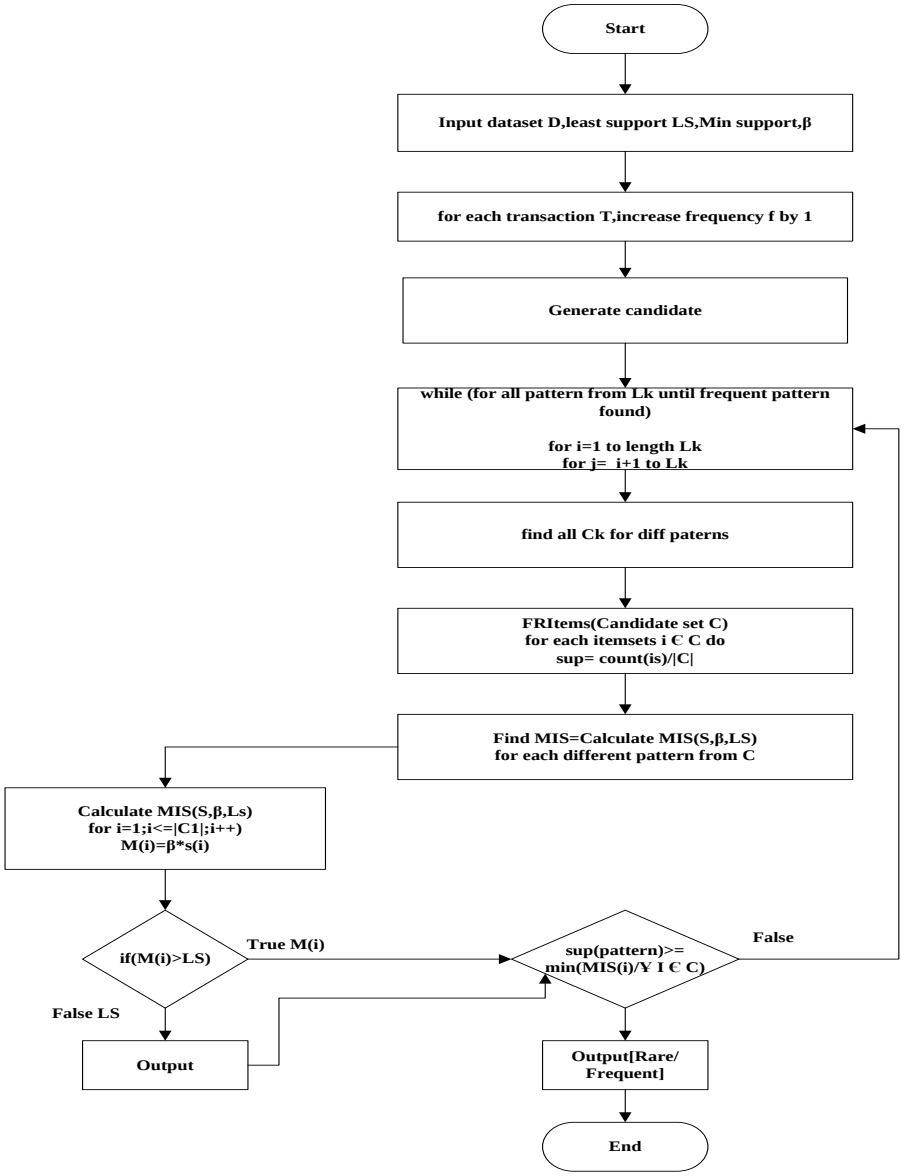


Figure 2:-Flowchart of the proposed system

Proposed Algorithm :-

Input:-Dataset D,least support,min support S,β.

Output:- Frequent and rare itmsets.

Step 1:-Represent items of transaction in binary form

According to table 2

For each transaction which is in binary form, increase the frequency as 1 for each item occurs in T

end for

Step 2-Generate candidate 1- itemset C1

FRItems(C1)

k=1

Step 3:-While (for all pattern from Lk untill frequent pattern found)

for i=1 to length(Lk)

for j=i+1 to length(Lk)

Find Ck+1 for (i+1) different patterns

Lk+1=FRItems(Ck+1)

K++

FRItems(Candidate set C)

for each itemsets i ∈ C do

sup= count(is)/|C|

end for

Find MIS=Calculate MIS(S,β,LS)

for each different pattern from C

if sup(pattern)>=) min(MIS(i)|∀i∈(C)

add pattern to L as frequent or rare item sets pattern

end for

return L

Step 4:- MIS(S,β,Ls)

for i=1;i<=|C1|;i++

M(i)=β*s(i)

if(M(i)>LS)

MIS(i)=M(i)

else

MIS(i)=LS

end if

end for

return MIS

Implementation:-

	β =0.7 LS=0.4	β =0.8 LS=0.4	β =0.9 LS=0.4
Frequent Items	51565	8865	757
Rare Items	7	3	1
Memory(MB)	136.8571	157.5855	32.39231
Time(MS)	34427	9394	3819

Table 7:- Table 4.1 : Comparison of MSApriori approach with Chess Dataset

	β =0.7 LS=0.4	β =0.8 LS=0.4	β =0.9 LS=0.4
Frequent Items	51565	8865	757
Rare Items	7	3	1
Memory(MB)	16.33111	12.5765	17.2591
Time(MS)	13315	5255	3504

Table 8 : Comparison of proposed SemiApriori approach with Chess Dataset

	β =0.2 LS=0.4	β =0.5 LS=0.4	β =0.6 LS=0.4
Frequent Items	505	203	86
Rare Items	505	172	41
Memory(MB)	102.357	22.7430	102.2845
Time(MS)	3436	2611	3297

Table 9 : Comparison of proposed SemiApriori approach with Chess Dataset

	β = 0.2 LS=0.4	β =0.5 LS=0.4	β = 0.6 LS=0.4
Frequent Items	505	203	86
Rare Items	505	172	41
Memory(MB)	86.3419	5.3825	5.3434
Time(MS)	3088	2359	2291

Table 10 : Comparison of proposed SemiApriori approach with Mushrooms Dataset

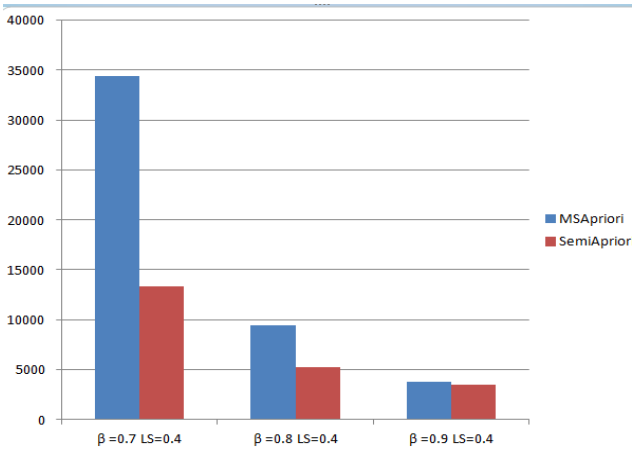


Figure 3 : Graphical representation of time comparison with Chess dataset

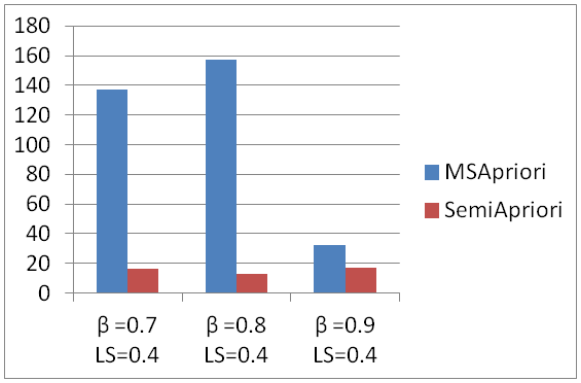


Figure 4 : Graphical representation of memory comparison with Chess dataset

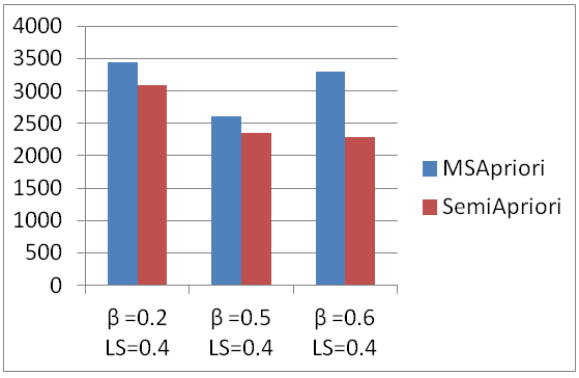


Figure 5 : Graphical representation of time comparison with Mushroom dataset

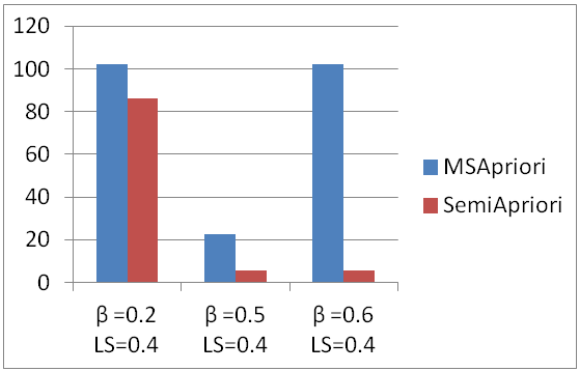


Figure 6 : Graphical representation of memory comparison with Mushroom dataset

CONCLUSION & FUTURE WORK:-

Non frequent patterns in customer purchasing behavior lead to identify new and unexploited business opportunities. Proposed approach avoids repeated scans, avoids generation of large number of candidate sets, minimize execution time. The proposed approach dynamically assigns appropriate minimum support to each item. As a result it happen so that frequent itemsets involving rare items can be extracted in a more efficient manner as compared to the existing approaches.

This SemiApriori algorithm successfully reduces time and memory and also gives rare and frequent itemisets, further it may be expanded that some tree structure based algorithm which can be developed which dynamically assigns MIS values and is efficient enough in reducing time and space complexity giving both frequent and rare itemsets.

REFERENCES:-

[1]. Sallam Osman Fageeri "A Semi-Apriori Algorithm for Discovering the Frequent Itemsets " 978-1-4799-0059-6/13 © 2014 IEEE
[2]. Jlolmo, Jrromero, Sventura "On the Use of Ant Programming for Mining Rare Association Rules ", World Congress on Nature and Biologically Inspired Computing (NaBIC),IEEE,pp. 220-225,2013
[3]. Luigi Troiano,Giacomo Scibelli,Cosimon Birtolo "A Fast Algorithm for Mining Rare Itemsets",9th International Conference on Intelligent Systems Design and Applications,IEEE,pp. 1149-1155,2009
[4]. Jutamas Tempaiboolkul "Mining Rare Association Rules in a Distributed Environment using Multiple Minimum Supports", Asian Institute of Technology, Thailand, IEEE,pp. 295-300,2013
[5]. R.Uday Kiran,P Krishna Reddy "An improved Multiple Minimum Support Based Approach to Mine Rare Association Rules",CIDM,IEEE,2009
[6]. R.Uday Kiran,P Krishna Reddy "An Efficient Approach to Mine Rare Association Rules Using Maximum Items Support Constraints", Verlag Berlin Heidelberg,Springer,pp. 84-95,2012
[7]. D. Braha, "Data mining for design and manufacturing: methods and applications: Kluwer academic publishers", 2001
[8]. Han and Kamber , "Data Mining:Concepts and Techniques " , Second Edition
[9]. Agrawal, R., Imielinski, T. & Swami, A. (1993), Mining association rules between sets of items in large databases,in P. Buneman & S. Jajodia, eds, 'Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data', pp. 207 – 216
[10]. http://www.thearling.com/text/dmwhite/dmwhite.htm(9th November 2014,10:23AM)
[11]. http://www.igi-global.com/chapter/rare-association-rule-mining/36896(2nd November 2014,4:38 PM)