An Improved Progressive Sampling based Approach for Association Rule Mining

S. S. Thakur
Department of Applied Mathematics
Jabalpur Engineering College,
Jabalpur (M.P.)

Shalini Zanzote Ninoria
Department of Mathematics and Computer Science
Rani Durgavati Vishwavidyalaya,
Jabalpur (M.P.)

ABSTRACT
Data Mining is the multistage process of extraction of useful information from the large database. Association rule mining is one of the important techniques of data mining in which relationships among the items present in the transactions are discovered. There are different algorithms are available in the field of data mining for association rule mining but most of them are time consuming hence the run time and memory overheads incurred is extremely high specially in the case of very large database. Sampling is one of the remarkable approach which can be used to speed up the process of association rule mining hence it is a approach to reduce the complexity of association rule mining technique to some extent but still consuming comparable time and memory. A progressive sampling based approach is a noval expert approach in the field of association rule mining to reduce the overheads of usual sampling based approaches. It is very effective in case of the large databases. In this paper, we have extended the Progressive sampling based approach presented by Umarani & Punithavalli,2009[22] and performed an extensive experimental analysis of the progressive sampling-based approach for the different Partitioned itemset 1/3,1/4,2/3,3/4 with the sample dataset also in addition the performance of this Improved Progressive Sampling Based Approach is evaluated with the Progressive sampling based approach by Umarani & Punithavalli,2009[22]. The experimental results illustrate the complexity of an algorithm in terms of run time as well as the memory utilization. Complete implementation has been done in Java Jdk 6.1. and MySQL5.0 on the Sample dataset CompPeriPurchase.

General Terms
Data Mining

Keywords
Association Rule Mining, Frequent Itemsets, Negative Border, Partitioned Itemsets.

1. INTRODUCTION
Today's world is an information era as we know the enhanced technologies require immense use of information which can be collected from the various sources and also can be exploit in various areas. For the past three decades, companies and organizations have stockpiled not only gigabytes but terabytes of data. This data presents a great untapped opportunity for knowledge discovery. Data mining emerged in 1990s and has a big impact in business, industry, and science [22]. Today, we have various types of information from the different source of information such as: information from business associations, information from scientific statistics, the satellite pictures, various text reports, information from military intelligence, medical researches, surveys etc. Only the information extraction is not sufficient to help in decision making. It is essential, to build up a powerful way for analysis of such data for the extraction of interesting knowledge that could help in decision-making. Data Mining, popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [11].

Association rule mining is a method of finding the relations between the variables in the database. The concept of Association Rules for discovering regularities between products in large databases has introduced by Rakesh Agrawal et al,1994[1,2].Association rule mining: "Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transactional databases, relational databases, and other information repositories. In example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk." So Association rule mining is the most valuable and well explored data mining technique, which is used by most of the organizations for decision making so that they improve the profit and enhance their performance in terms of sales and good product quality.

The traditional Classical association rule mining algorithms much more time consuming process as required more number of passes to accomplish the task. Hence sampling approach can be used to solve this scalability problem.

In the earlier research, Umarani & Punithavalli, 2009[22] presented an efficient progressive sampling-based approach for mining of association rules from databases. The approach selects an initial sample of the database. Initially, frequent itemsets are mined from the initial sample using Apriori algorithm. Subsequently, the negative border is computed and the itemsets in the negative border are sorted based on their support level. The midpoint itemset is scanned on the remaining set of records in the input database to find the support level. If the support of the midpoint itemset is greater than the user specified support, the chosen sample size is progressively increased. This procedure is repeated until an optimal sample size is obtained and then, association rules are mined from the optimal sample. Finally, the support of the midpoint itemset is analyzed with the different percentage of databases. In this paper, we have improved and validate the earlier approach with different Partitioned Itemsets instead of the midpoint itemset in the negative border itemsets. Furthermore; our improved approach is evaluated with the previous algorithm in terms of memory and time complexity.

The rest of the paper is organized as follows: Section II gives review of related researches on sampling-based association rule mining. Section III presents the progressive sampling-based approach for association rule mining. Section IV presents the improved progressive sampling-based approach for association rule mining. The experimental results and analysis is given in Section V Conclusion and Future Enhancement is summed up in Section VI.

2. LITERATURE REVIEW

The literature presents study with numerous algorithms for Association Rule Mining, among which Apriori has been the most renowned mainly owing to its efficiency in knowledge discovery [1,2]. The execution of traditional association rule mining algorithms that necessitate multiple number of passes over the complete database, can consume hours or even days, and in the future, this problem will only become even worse. Hence recently, to reduce this adverse effect, researches have intended to develop efficient approaches that reduce the I/O and computational requirement of the ARM techniques [21,22]. As concern with the large data base size the sampling approaches also gives challenges to find the frequent item sets. So how to determine a fitting sample size is a vital means for the success of the sampling techniques. Therefore as to determine the optimal sample size quickly, researchers have lately turned to progressive sampling. The objective of progressive sampling is to start with tiny samples and increase them progressively given that model accuracy improves adequately [18]. Several researches are available in the literature for sampling-based association rule mining. A brief review of some of the significant researches is presented . (Toivonen et al, 1996) [19] Presented an Association rule mining algorithm using sampling. The approach can be divided into 2 phases. In phase I, a sample of the database is obtained and all associations are found. These results are then validated against the entire database. To maximize the effectiveness of the overall approach the author makes use of lowered minimum support on the sample.(Mohammed et al, 1997)[15] reviews and proposed that random sampling of transactions in the database is an effective way for finding association rules. They have the following contributions i. Sampling can reduce i/o cost by drastically shrinking the number of transactions to be considered and ii. Sampling can provide greater accuracy with respect to the association rules. They have shown that sampling can speed up the mining process by more than a order of magnitude. The validity of the sample is determined by two characteristics namely the size and quality of the sample. The quality, in the context of statistical sampling techniques, refers to whether the sample captures the characteristics of the database. The highest quality sample would be an exact miniature of the database; it would preserve the distributions of individual variables and the relationships among variables [20].The quality of the sample for association rule mining can be improved by considering factors like transaction length and transaction frequency [4]. (Venkatesan T et al, 2009) [25] have presented a comprehensive theoretical analysis of the sampling technique for the association rule mining problem. The sampling based technique was used to solve frequent itemset mining and association rule mining problems using a sample whose size is independent of both the number of items and the number of transactions. Thus, the possibility of speeding up the entire process of association rule mining for huge databases by working with a small sample while retaining any desired degree of accuracy was established. The proposed approach is likely to give considerable reduction in computational time with some cost to accuracy (optimality between accuracy and time).

In most general cases, if the rules mined from a sample are unsatisfactory, the sample size is increased, and the mining step is executed again. An iterative procedure is followed in increasing the size of the sample until interesting rules are found. In the above scenarios, how to determine a fitting sample size is a vital means for the success of the sampling technique. So as to determine the optimal sample size quickly, researchers have lately turned to progressive sampling. The objective of progressive sampling is to start with tiny samples and increase them progressively given that model accuracy improves adequately [22].

Sampling can speed up the mining process. Apriori is a renowned algorithm for Association Rule Mining but most of the prior works on sampling have concentrated on speeding up the phase by running a frequent itemset mining algorithm only on a small sample of the database. Next Reducing the Number of Passes. The disadvantage of Apriori algorithm made the researchers to think about new techniques to mine frequent patterns. Sampling is a very powerful data reduction technique which can be utilized to various problems of data mining. Thus the different methods discussed above are provides that the Association Rule Mining algorithm such as sampling can be used for performance enhancement.

Hence overall a sample based approach is an optimal solution for effective mining of association rules from a large database has been proposed. In this method two basic operations performed first an initialize sample of certain databases. Based on the frequent itemsets generated from the initial sample, the previous approach computes a negative border secondly, then the sample size was either progressively increased or association rule are mined by regarding it as an optimal sample [22].The Negative border get sorted and portioned into two equal parts in the previous approach instead in proposed improved approach the portioning is done in different slots in the list.

3. PROGRESSIVE SAMPLING BASED APPROACH

As we all know that the frequent itemset mining is the most important step in the entire data mining process. Apriori had flashed a lot of interest in the data mining community. And afterwards many researchers have proposed different ways to improve Apriori and some are still working on it.. Yet, there are many more researchers who are continuing to work on frequent itemset mining. The Apriori Algorithm is a powerful algorithm for mining frequent itemsets.

Key Concepts: Frequent Itemsets: The sets of item which has minimum Support (denoted by L). As Per (Agarwal et al, 1994)[2] the task of association rule mining has a great deal of attention. Today the mining of such rules is still one of the popular pattern-discovery methods in Knowledge Discovery in Databases. ARM techniques have found their widespread application in marketing and retail communities in addition to many other diverse fields [6,13].

Apriori is a powerful algorithm for mining frequent itemsets for association rules (Agarwal et al, 1994)[1]. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore k+1 itemsets.

The Advantage of the Apriori algorithm is perfect pruning of infrequent candidate item sets (with infrequent subsets). While on the other hand, the disadvantage of Apriori algorithm is that can require a lot of memory (since all frequent itemsets are represented) and support counting takes very long for large transactions [25].

Following is the Apriori algorithm:

$I_1 = \{large\ 1\text{-}itemsets\};$
$For (k = 2; I_{k-1} \neq \phi; k++)\ do\ begin$
$C_K = apriori\text{-}gen(I_k - 1); //New\ Candidates$
$For\ all\ transactions\ T \in D\ do\ begin$
$C_T = subset(C_k, T); //Candidates\ contained\ in\ T$
$for\ all\ candidates\ c \in C_T\ do$
$c.count++;$
end
end
$I_k = \{c \in C_k, c.count \geq minsup\}$
end
$Answer = U_k I_k;$

3.1 Progressive Sampling-Based Approach for Effective Association Rule Mining:

The main aim of our earlier research is to determine the association rules from the original database and at the same time, it should be capable of providing less computation time with accuracy. To facilitate, the approach used the progressive sampling-based technique for discovering the optimal sample size with the help of negative border. Negtive border itemsets can be defined as those itemsets that do not satisfy the minimum support. [24] The important steps used for effective mining of association rules, which is based on the progressive sampling-based approach are as follows:

Following are the steps involved in this method:

1. Selection of initial sample S_i (systematic sampling) of size 'n' based on the temporal characteristics of Database D.
2. Generation of frequent item sets and negative border using Apriori algorithm.
3. Sorting of the negative border based on the support level of itemsets.
4. Selection of the mid itemset in the sorted negative border.
5. Single scan on the left behind records of the database D.
6. If the support computed for the midpoint itemset is less than the specified support, the selected sample is optimal and association rules are mined from it. Else, the sample size ' n ' is increased and steps 2-5 are performed progressively until an optimal sample size is achieved.

The core idea behind the proposed approach is that, progressive sampling, based on negative border will enable in determining an optimal sample size for effective mining of association rules [23]

Below given is the pseudo code of this previous progressive sampling based approach:

Input: Transaction database D, Size of initial sample in %,
increment of sample size in %, minsupp, minconf
Output: Optiml Sample Size, Association Rules
Assumption
TData → Transaxtion Data Items
S_i → Sample Selected from TDFata
Size → Sample Size
NB → Negtive Border
NB' → Sorted Negative Border
Miditemset → Miditemset
Min Supp → Minimum Support Threshold
Min Conf → Minimum confidence Threshold
F_{ki} → Frequent Itemsets
Rules → Association Rules

```
Begin
      TData = {t : t ∈ D};
Z:
Si = φ;
Si = getTempSample(TData,Size);
[Fki, NB] = Apriori(Si, min supp);
NB'= sort(NB)desc;//basedon minsupp
miditem = NB'[\NB'\/2];
If(support(miditem < minsupp)
  Rules = genrules(Fki);
Else
   Size = Size + inc;
   GotoZ;
EndIf
```

4. IMPROVED PROGRESSIVE SAMPLING BASED APPROACH

The developed approaches adopt the beliefs of Apriori with Progressive Sampling Based Approach with some modifications in order to reduce the time execution and memory utilization of the algorithm. Here we have extended the algorithm of sampling based approach for association rule mining[21,22]. In our proposed improved approach we have taken the partitioning of the dataset instead of two equal partitions, here the approach portioned the dataset into the parts of 3 or more. We have taken the partition of itemsets on the 2/3, 1/4, 1/3 & 3/4 parts. The Sample is passed through the Apriori algorithm and then it finds the negative border elements and gets sorted. The sorted dataset than partitioned into the various sections and then the frequent itemsets will generate the association rules for data mining.

Input: Transaction database D, Size of initial sample in %,
increment of sample size in %, minsupp, minconf
Output: Optiml Sample Size, Association Rules
Assumption
TData → Transaxtion Data Items
S_i → Sample Selected from TDFata
Size → Sample Size
NB → Negtive Border
NB' → Sorted Negative Border
Partitionditem → Partitioning Itemset in the Sorted Negative Border
Min Supp → Minimum Support Threshold
Min Conf → Minimum confidence Threshold
F_{ki} → Frequent Itemsets
Rules → Association Rules

4.1 Improved Proposed Approach Algorithm(1/4th Partitioned Itemset)

```
Begin
      TData = {t : t ∈ D};
Z:
S_i = φ;
S_i = getTempSample(TData,Size);
[F_ki, NB] = Apriori(S_i, min supp);
NB' = sort(NB) desc; //based on minsupp
Partitioneditem = NB'[\NB'\/4];
If(support(miditem < minsupp)
  Rules = genrules(F_ki);
Else
   Size = Size + inc;
   GotoZ;
EndIf
```

4.2 Improved Proposed Approach Algorithm (3/4th Partitioned Itemset)

```
Begin
      TData = {t : t ∈ D};
Z:
S_i = φ;
S_i = getTempSample(TData,Size);
[F_ki, NB] = Apriori(S_i, min supp);
NB' = sort(NB) desc; //based on minsupp
Partitioneditem = 3*NB'[\NB'\/4];
If(support(miditem < minsupp)
  Rules = genrules(F_ki);
Else
   Size = Size + inc;
   GotoZ;
EndIf
```

4.3 Improved Proposed Approach Algorithm (1/3rd Partitioned Itemset)

```
Begin
      TData = {t : t ∈ D};
Z:
S_i = φ;
S_i = getTempSample(TData,Size);
[F_ki, NB] = Apriori(S_i, min supp);
NB' = sort(NB) desc; //based on minsupp
Partitioneditem = NB'[\NB'\/3];
If(support(miditem < minsupp)
  Rules = genrules(F_ki);
Else
   Size = Size + inc;
   GotoZ;
EndIf
```

4.4 Improved Proposed Approach Algorithm (2/3rd Partitioned Itemset)

```
Begin
      TData = {t : t ∈ D};
Z:
S_i = φ;
S_i = getTempSample(TData,Size);
[F_ki, NB] = Apriori(S_i, min supp);
NB' = sort(NB) desc; //based on minsupp
Partitioneditem = 2*NB'[\NB'\/3];
If(support(miditem < minsupp)
  Rules = genrules(F_ki);
Else
   Size = Size + inc;
   GotoZ;
EndIf
```

The procedure we have followed here in this proposed approach is –

1. Firstly we take the Sample Dataset called S_i.
2. Apply Apriori Algorithm on $S_{i,\ for}$ Association Rule Mining.Assoc iation rule mining can be done in two different phases, namely (i) frequent itemsets generation (ii) mining of association rules.
3. Generate the Negative Border itemsets.
4. Sorting of the negative border based on the support level of itemsets.
5. Selection of the Partitioning itemset in the sorted negative border.
6. Approaches are implemented in 4 different ways – in [A] the Partitioned Itemset in sorted negative border is taken as $1/4^{th}$, in [B] the Partitioned Itemset in sorted negative border is taken as $3/4^{th}$, in [C] the Partitioned Itemset in sorted negative border is taken as $1/3^{rd}$, in [A] the Partitioned Itemset in sorted negative border is taken as $2/3^{rd}$.
7. After Partitioning the sorted negative border scaning of the remaining datasets will be done on the Dataset.
8. If the support computed for the partitioned t itemset is less than the specified support, the selected sample S_i is optimal and association rules are mined from it. Else no rules will be generated and steps 2-5 are performed progressively and size of the sample is increased accordingly.

5. EXPERIMENTAL RESULTS AND ANALYSIS

This section illustrates the experimental results and extensive analysis of Improved progressive sampling-based approach in mining of association rules. The Improved progressive sampling-based approach is implemented in Java (jdk 1.6), NetBeans IDE 7.0.

The implementation has done on the dataset for CompPeriPurchase. These implementations have been done with sample datasets with varying of transactions, with

varying number of Partitioning Itemset and varying Support and Confidence. Below shows the number of distinct transactions and distinct items contained in the sample. **CompPeriPurchase**- The **CompPeriPurchase** dataset (282 bytes):

File Edit View Query Script Tools Window MySQL Enterprise Help

Go back Next Refresh

```
SELECT * FROM `abc`.`compperipurchase`
```

Resultset 1

pid	pname
1	COMPUTER
2	CAMERA
3	SPEAKER
4	CELLPHONE
5	PRINTER
6	MEMORYCARD
7	MP3PLAYER
8	PHOTOPRINTER
9	SCANNER
10	LIGHTPEN
11	PENDRIVE
12	MICROPHONE
13	WEBCAM
14	LCDMONITOR
15	LEDMONITOR
16	OPTICALMOUSE
17	USBMOUSE
18	UPS

Table1: CompPeriPurchase

Number of Instances = 18 Number of Attributes=Univarient No. of Transactions =37

Here we have implemented the experiment for the different confidence values with the threshold value of 35%, 40%, 50%, and 60% respectively. For each approach like 1/4, 3/4, 1/3, 2/3 respectively. The graph shows the growth for the different confidence values with the threshold value of 35%, 40%, 50%, and 60% respectively.

All figures given below shows the confidence growth which is upgraded as the support count values remain in increasing order. The complete growth of the confidence and support count is dependent on the number of transactions. If we get the confidence in increasing order than definitely we can achieve more association rules and we can lift more frequent item sets pair for the best association rule generation.

5.1 Implementation of an Improved Progressive sampling based algorithm with (1/4th Partitioning itemset):

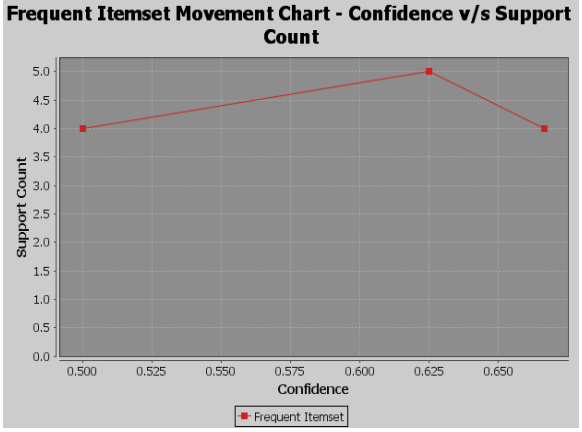

Fig.1 Movement of Frequent Itemset for diff. confidence and corresponding support count (1/4th partitioning itemset).

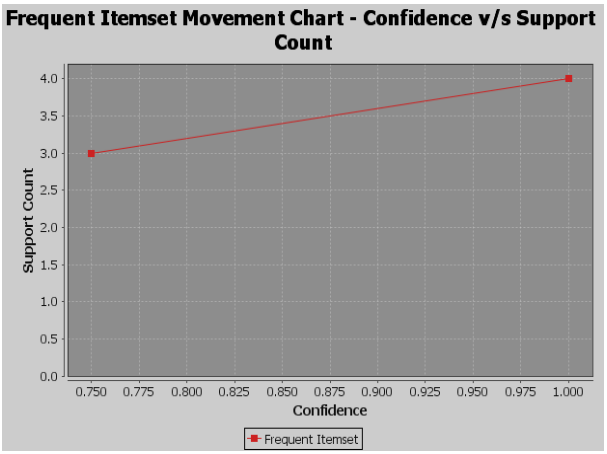

Fig.2 Movement of Frequent Itemset for diff. confidence and corresponding support count (1/4^{t} Partitioning itemset).

5.2 Implementation of an Improved Progressive sampling based algorithm with (3/4th Partitioning itemset)

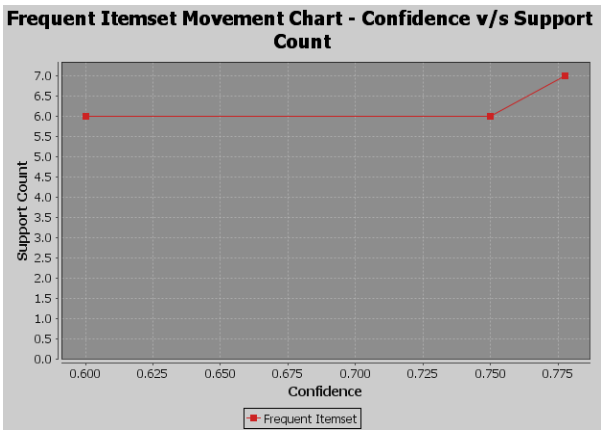

Fig.3 Movement of Frequent Itemset for diff. confidence and corresponding support count (3/4thPartitioning itemset)

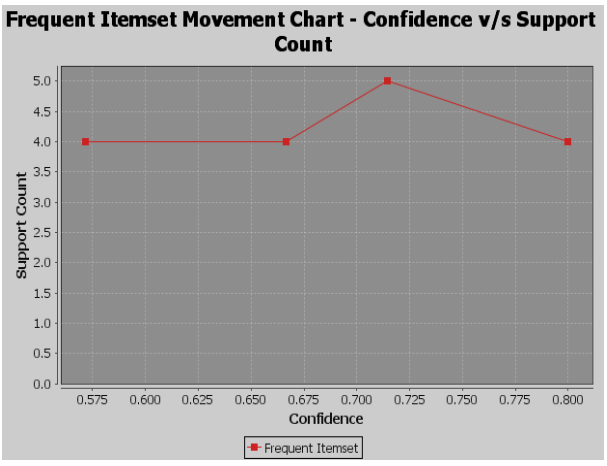

Fig.4 Movement of Frequent Itemset for diff. confidence and corresponding support count (3/4th Partitioning itemset).

5.3 Implementation of an Improved Progressive sampling based algorithm with (1/3rd Partitioning itemset)

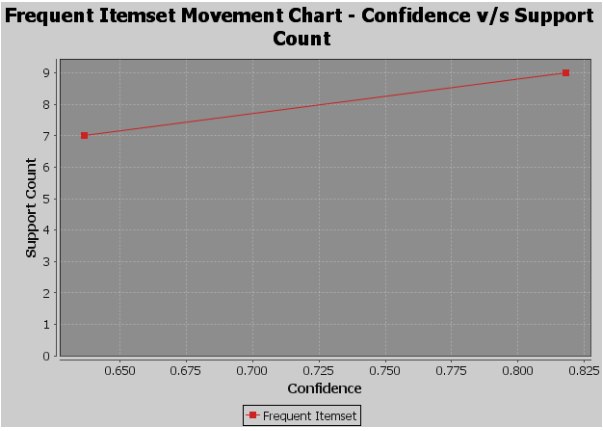

Fig.5 Movement of Frequent Itemset for diff.confidence and corresponding support count(1/3thPartitioning itemset)

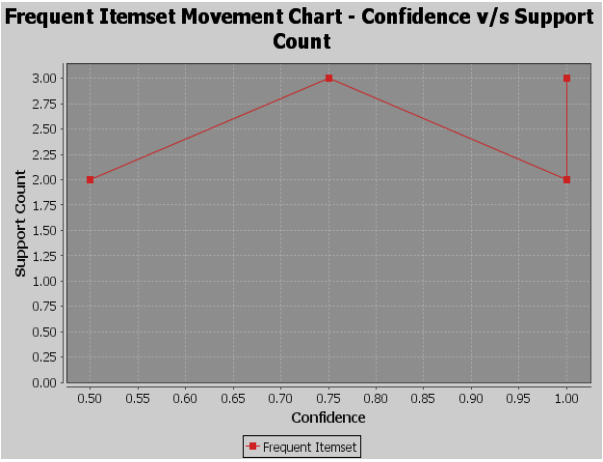

Fig.6 Movement of Frequent Itemset for diff.confidence and corresponding support count (1/3thPartitioning itemset).

5.4 Implementation of an Improved Progressive sampling based algorithm with (2/3rd Partitioning itemset)

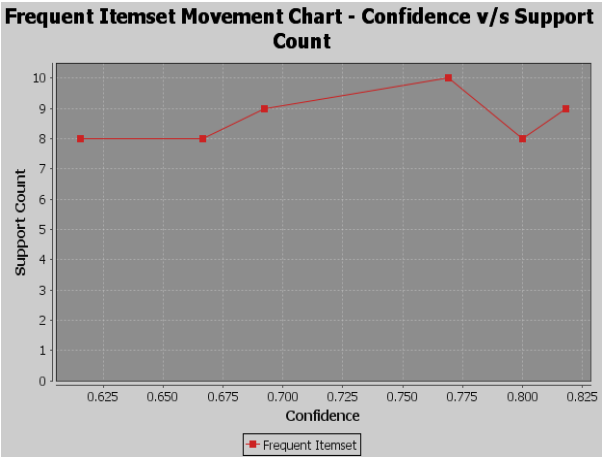

Fig.7 Movement of Frequent Itemset for diff.confidence and corresponding support count (2/3thPartitioning itemset)

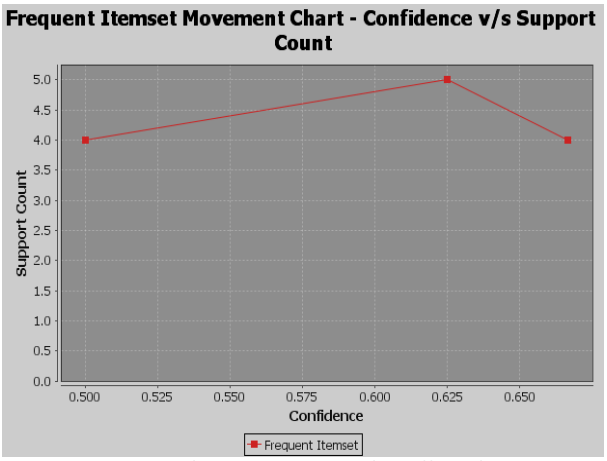

Fig.8 Movement of Frequent Itemset for diff. confidence and corresponding support count (2/3th Partitioning itemset).

Hence we have analyzed in this section the results obtained from the experimentation on the Improved Progressive Sampling Based Approach for Association Rule Mining. The experiments are conducted in order to assess the practical feasibility of using different support thresholds from Lower Thresholds to Higher Thresholds for finding frequent itemsets and generate association rules.

This implementation has been compared with previous Sampling Based Approach for Association Rule Mining [21].

For evaluating the performance of the Improved Progressive Sampling-based Approach, we have chosen two parameters such as, i) Run Time ii) Memory Utilization.

5.5 Runtime

In computer science, the analysis of algorithms is for the determination of the amount of resources (such as time and storage) necessary to execute them. Here, runtimes of the Improved Progressive Sampling Based Algorithms are compared with the Previous Progressive Sampling Based Algorithm for various threshold values on the sample datasets. Results are shown in below graph.

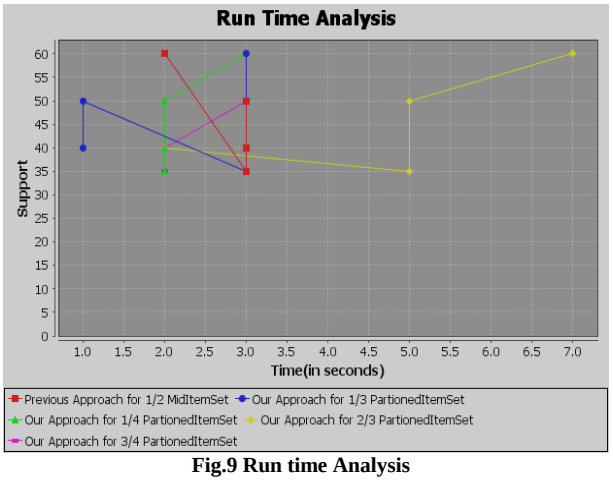

Fig.9 Run time Analysis

It can be observed in the above figure that the proposed Improved Progressive Sampling Based Algorithms enhanced than previous algorithm for the various thresholds including both lower and higher thresholds, on sample datasets. We can observe easily that our four algorithms act differently for various threshold values.

For example the run time mentioned for the lower thresholds i.e. the 35% and 40% the previous approach acquire 3 seconds while our Proposed Improved Progressive Sampling Based Approach(3/4th PartitionedItemset) acquire only 2 seconds run time which is comparative to the previous approach, and shows run time reduction of complete 1 second. Similarly our proposed approaches with 1/4th and 1/3rd PartitionedItemset also acquire comparatively lesser time for lower threshold values i.e. 3seconds and 1 second respectively which is again better than the previous approach.

For higher thresholds i.e. 50% and 60% in our approaches (3/4th,1/4th & 1/3rd) PartitionedItemsets it acquire 3 second,1 second and 3 second which is not comparable with the previous approach. Also one of our proposed approach for 2/3rd PartitionedItemset the time required is 5,2,5,7 seconds for lower & higher thresholds respectively which is not upto the mark with the previous approach. This is because the compared algorithms are all sensitive to transaction length.

Hence we can summarize that our Proposed Improved Progressive Sampling Based Approaches works efficiently with lower threshold values than the higher threshold values in terms of run time.

5.6 Memory Usage

Memory complexity is the size of work memory used by an algorithm. The memory usage of the algorithms was also compared when varying the thresholds from lower to higher, for sample datasets.

Results are shown in below Figure.

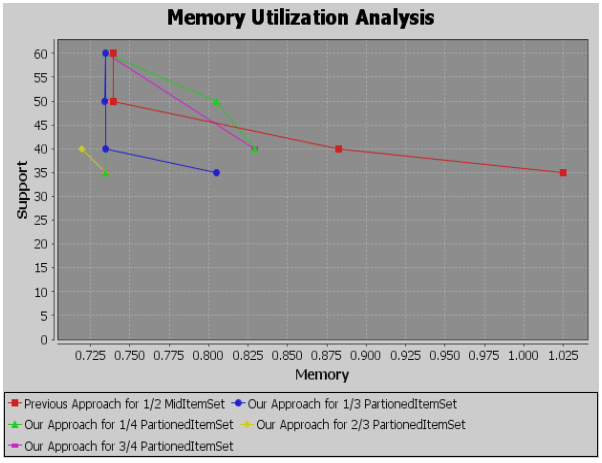

Fig.10 Memory Utilization Analysis

It can be observed that the proposed algorithm outperforms in terms of memory usage for various threshold values from lower to higher.

Here we can see from the Figure the memory utilization of previous approach is1.02508544921875, 0.882301330566406,0.739547729492188,0.739364624023437and for our Proposed Improved Progressive Sampling Based Approaches (1/4th ,1/3rd ,3/4th and 2/3$^{rd)}$ Partitioned Itemsets lies only between 0.719320258789063 to 0.829290258789062 for all lower and higher threshold values i.e. 35%,40%,50%& 60%.

Hence we can summarize that our Proposed Improved Progressive Sampling Based Approaches works efficiently with lower threshold values as well as with higher threshold values in terms of memory usage.

6. CONCLUSION AND FUTURE ENHANCEMENT

In this paper we have discussed two approaches (Progressive Sampling Based Approach[22] and our proposed Improved Progressive sampling Based Approaches) for association rule mining in the context of negative border partitioning. We have discussed about the frequent itemsets generation as well as rule generation. Rule generation is very simple as compared to frequent itemsets mining and it requires very less time. Extensive experiments have been performed to test the performance of these approaches over our sample dataset. The Improved Progressive Sampling based approaches gives better complexity than the previously progressive sampling based approach in terms of Memory Usage for lower threshold as well as for higher thresholds. Also our Proposed Improved Progressive Sampling Based Approaches works efficiently with lower threshold values than the higher threshold values in terms of Run Time.

Sr. No.	Algorithms (Previous and Proposed)	Support (Threshold %)	Memory Utilization in bytes	Run Time in Seconds
1	Previous Progressive Sampling Based Approach (MidItemset)	35%	1.025085449	3
		40%	0.882301331	3
		50%	0.739547729	3
		60%	0.739364624	2
2	Proposed Improved Progressive Sampling Based Approach (1/4th PartitionedItemset)	35%	0.734390259	3
		40%	0.829290259	1
		50%	0.804931455	1
		60%	0.734390259	3
3	Proposed Improved Progressive Sampling Based Approach (3/4th	35%	0.734334339	2
		40%	0.829390256	2
		50%	0.734390259	3
		60%	0.734390259	3
4	Proposed Improved Progressive Sampling Based Approach (1/3rd	35%	0.80493927	3
		40%	0.734390783	1
		50%	0.734127159	1
		60%	0.734390259	3
5	Proposed Improved Progressive Sampling Based Approach (2/3rd	35%	0.734345446	5
		40%	0.719320259	2
		50%	0.734413147	5
		60%	0.734390259	7

Table 2 Conclusion Table

Now we can conclude that the usage of memory and time for all of our proposed approaches as follows,

1. Proposed Improved Progressive Sampling Based Approach (1/4th PartitionedItemset): In this approach the memory utilization is reduced for both the higher and lower threshold values. Also the run time is reduced for lower threshold values but not for higher threshold value.

2. Proposed Improved Progressive Sampling Based Approach (3/4th PartitionedItemset): In this approach Memory Utilization is reduced for both higher and lower threshold values and run time is also reduced except for the higher threshold value.

3. Proposed Improved Progressive Sampling Based

Approach (1/3rd PartitionedItemset): In this approach Memory Utilization is reduced for both higher and lower threshold values and run time is also reduced except for the higher threshold value.

4. Proposed Improved Progressive Sampling Based Approach (2/3rd PartitionedItemset): In this approach Memory Utilization is reduced for both higher and lower threshold values but the run time increased rapidly compare to the previous approach. Hence in concern to the run time reduction we do not recommend this approach.

So overall we can conclude that our all four proposed algorithms are outperforms in terms of Memory Utilization for lower as well as higher threshold values. Similarly for run time reduction our proposed algorithms performs enhanced for lower threshold values except the Proposed Improved Progressive Sampling Based Approach (2/3rd PartitionedItemset).

Hence we recommend our proposed Improved Progressive Sampling Based Approach instead of previous approach for better Memory Utilization. Also we strongly recommend our Proposed Improved Progressive Sampling Based Approach (1/3rd PartitionedItemset) to acquire both reduced memory usage for higher and lower thresholds as well as reduced run time for lower thresholds.

6.1 Future Enhancement

Some of the future enhancements of the paper are presented below:

- The work presented in the paper can be extended for multi-level association rule mining.
- The work can be enhanced to generate multi-dimensional association rules.
- A tool for generating association rules can be developed. This tool can choose the approach for frequent itemsets mining according to the properties of the dataset to be mined.

7. REFERENCES

[1] Agarwal, R. and Srikant, R. "Fast algorithms for mining association rules", In the Proceedings of 20th Int'l Conf. Very large Data Bases, pp.487-499,1994.

[2] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*. p. 207

[3] Atul Palandurkar, "NetBeans IDE How-to", PACKT Publishing, UK (2013).

[4] B. Mobasher, N. Jain, E.H. Han, and J. Srivastava, "Web Mining: Pattern Discovery from World Wide Web Transactions" Department of Computer Science, University of Minnesota, Technical Report TR96-050, (March, 1996).

[5] Basel et al, "a new sampling technique for Association rule mining", Journal of information science ,June 2009,vol 35, pp 358-376.

[6] Bodon, F. "A Fast Apriori Implementation", In the Proceedings of the IEEE ICDM Workshop on Frequent Item set Mining Implementations, Vol.90, Melbourne, 2003.

[7] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006

[8] Farah Hanna AL-Zawaidah , Marwan AL-Abed Abu-Zanona, Yosef Hasan Jbara,"An Improved Algorithm for Mining Association Rules in Large Databases"(WCSIT) ISSN: 2221-0741 Vol. 1, No. 7, 311-316, 2011

[9] G.K.Gupta, "Introduction to Data Mining with Case Studies"Prentice-Hall of India Pvt.Ltd.New Delhi,India(2006)

[10] Herbert Schildt, "Java : The Complete Reference, Seventh Edition"Tata Mc Grow Hills Publishing Company Ltd. ,New Delhi.

[11] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[12] Jiawei Han and Micheline Kamber "Data Mining – concept and techniques" Morgan Kaufmann Elsevier Science India (2002)

[13] Jiming, L. Yiuiming, C. and Hujun, Y. "Intelligent data engineering and automated learning", In the Proceedings of the 4th International Conference IDEAL Springer, 2003.

[14] Maojo V., Sanandres J., A Survey of Data Mining Techniques, LECT NOTES COMPUT SC, 2000, 1933, 77-92.

[15] Mohammed Javeed,Zaki,Srinivasan Parthasarathy,Wei Li,Mitsunori Ogihara,"Evaluation of Sampling for data mining of Association rules", proc Intn'l workshop research issues in data engineering 1997.

[16] Sotiris, K. and Dimitris, K. "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol. 32, No.1, pp.71-82, 2006.

[17] Srikant, R., Vu, Q. and Agarwal, R. "Mining association rules with item constraints", In the proceedings of 3rd Intl Conf on Knowledge Discovery and Data Mining, 1997.

[18] Srinivasan Parthasarathy,"Efficient Progressive Sampling for Association Rules", in Proceeding of the 2002 IEEE International Conference on Data Mining,pp.354,2002.

[19] Toivonen.H.(1996),Sampling large databases for association rules in "The VLDB Journal" pp.134-135.

[20] Tsau, Y.L. "Sampling in association rule mining", International Conference on Data mining and knowledge discovery: Theory, Tools and Technology VI, Vol.5433, pp. 161-167, 2004.

[21] Umarani, V. and Punithavalli, M. "A Novel Progressive Sampling based Approach for Effective Mining of Association Rules", International Journal of Computer Science and Research, Vol. 10 Issue 11(2010).

[22] Umarani, V. and Punithavalli, M. "Developing Novel and Effective Approach For Association Rule Mining Using Progressove Sampling", International Conferecne on Computer and Electrical Engineering ,IEEE,(2009).

[23] Umarani, V. and Punithavalli, M. "Sampling Based Association Rule Mining–A recent overview", International Journal of Computer Science and Engineering", Vol. 2 Issue 2 (2010).

[24] V.Umarani, M.Punithavalli," On developing an effectual progressive sampling based approach for Association Rule Discovery", In the proceedings of 2nd IEEE ICIME Int'l conference on Information and Data Management", Chengdu,China(2010)

[25] Venkatesan, T. C, Vinayaka, P. and Yogish, S. "Analysis of sampling techniques for Association Rule Mining.", In the Proceedings of the 12th International Conference on Database Theory, Vol.361, pp. 276-283, 2009.

[26] Vikram Vaswani, "MySQL : The Complete Reference,"Tata Mc Grow Hills Publishing Company Ltd., New Delhi.