iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48416

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Discovering Extraordinary items from BinaryEnhanced Database using APRIORI-BEAP Algorithm

R. Savitha[1], Dr. V. Baby Deepa[2]
[1, 2]*Research Scholar, Government Arts College (autonomous), Karur*

***Abstract:** An Item set is an instance format that is hidden beneath the raw data and after mining the database is created to help users make certain decisions to improve their business needs. Most of the previous studiesare carried out in extracting frequently occurring items or patterns and rare patterns are often ignored. Rare patterns are very useful in many areas like fraud detection, underselling products, intrusion attacks on networks, and these rare patterns indicate the user with some interesting detils to overcome and avoidthose problems. This paper proposes a new algorithm, BEAP algorithm, which uses binary representation and detects rare items without compromising speed and reduces memory consumption. From the experimental evaluation the proposed algorithm BEAP has been proved to be more effective and efficientthan the existing algorithms.*

I. INTRODUCTION

Due to the rise and evolution of the Internet, the amount of data available to businesses and individuals is increasing. Extracting useful data patterns (knowledge) from garbage data is called data mining and these extracted patterns are decoded and fed to some other process to interpret them to the common man. Slicedpatterns can be in textual or graphical form and can be extracted from numerous databases such as sequential, transactional and timestamp databases. Depending on the shape, the type of mine is used and used. Regular pattern mining approach is designed to find frequent objects in the data and this FI usually reveals the symmetries in the data. More concisely, they support items that occur frequently in the database, but rare items that do not occur frequently, but are of interest in many areas such as medical diagnosis and computer network attacks. So rarities are often referred to as abnormalities or anomalies inuser purchasing behavior [1] and indicate dubious security of the system [2]. Since sparse method miningis very difficult to make decisions, mining algorithms require some unique care and custom tailor solutions. Rare itemset mining is a new area and problems in the operational process have not yet been researched and analyzed.

II. BASIC CONCEPTS

Consider a1,a2,a3,…. ak} and TID = {T1,T2,T3.. TK}. A subset of Y of I is |Y| If denoted as itemset = k,then Y is a k-itemset. For example, a1a2 is a set of two items of I. Items {a1, a2, a3} are identified by the unique identifier {T1, T2, T3}.. Support is the actual measure or number (frequency) of items in a transaction. If the given number of items {a1,a2} is greater than or equal to the user-supplied support count value, the items are considered frequent or rare or infrequent.
A sample dataset is given in Table1.set of sample database is shown in figure 1

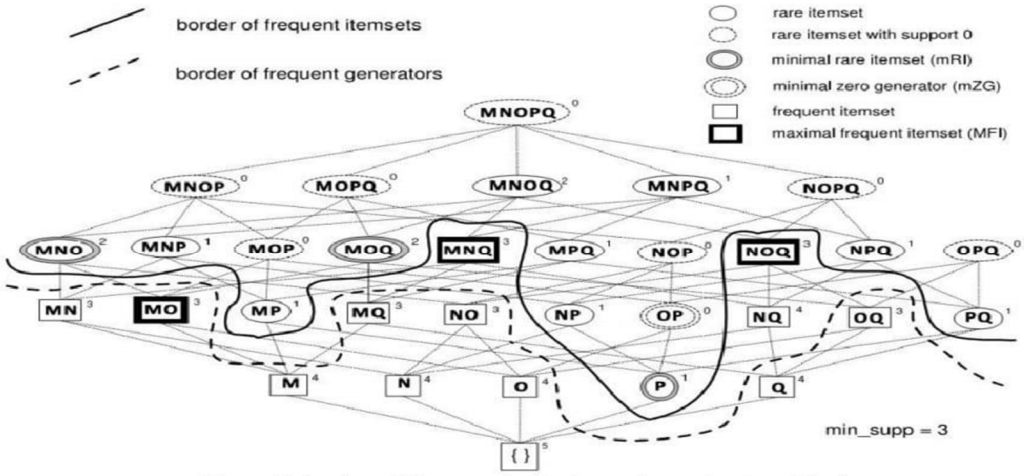


Figure 1: lattice of the sample database shown in the table 1

TID	ITEMS
1	M , N , P , Q
2	M , O
3	M , N , O , Q
4	N , O , Q
5	M , N , O , Q

Table 1: Sample database

III. OBJECTIVE OF THIS PAPER

The main objective of this thesis is to develop a new algorithm called "Binary Enhanced APriori - BEAP Algorithm". ” by combining binary representation and finding rare items from large databases. The secondary objectives are explained below,

1) To make a detailed comparison with existing algorithms in terms of running time and memoryconsumption using some standardized databases.

2) To analyze and study the current algorithms dealing with rare items mining using Aprioriapproach.

IV. PROBLEM BACKGROUND

Although frequent item mining has evolved to a great extent, finding frequent and rare items from a database is still a complex task because it deals with many calculations and functions related to support counting and pruning. Non-promised goods. The primary problem is generating a large number of frequent items, and this makes the algorithm inefficient. If the minimum support value is set to a very lowvalue, the amount of candidates generated will be high and if the minimum support value is set to a high value, rare items will be missed. This problem is fixed by introducing binary representation in the proposed work and starting to find rare items from the combination of unique items, and the proposed algorithm only handles 2U combinations, i.e. U are unique items.

V. FIND UNIQUE OBJECT FROM DB

Individual items from the sample database shown in Table 1 are initially found using the FindUniprocedure shown in Figure 2.

PROCEDURE FindUni (Database D)

ITEM	T1	T2	T3	T4	T5
M	1	1	1	0	1

INPUT : Database D
OUTPUT : Unique Items
BEGIN :

1. Load the input database D into the procedure FindUni
2. Initialize an array G[]=ϕ
3. Find the total number of rows in the DB —› T
4. For each row ri R
5. For each Item Ii ri
6. If [Ii Not IN G[ ]] do begin
7. G[ ] Ii// store unique items
8. Else
9. Count Ii value Ii counter // count unique items
10. Close IF
11. Close For
12. Close For

Return G[ ] with count
END

Figure 2: Pseudo code of the procedure to find the unique items.

Figure 2 shows the process of finding individual items from the database. The output of the process isshown in the following table 2.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

ITEMS	FREQUENCY
M	4
N	4
O	4
P	1
Q	4

Table 2: Unique items discovered

VI. REPRESENT THE DATA IN BINARY

The database is reconstructed in binary form, i.e. if there is an item in the row identifier, that item is marked as "1". A row identifier marked with "0" does not exist. For example, the unique item {M} is inTID's {T1, T2, T3, T5} and is denoted as , the process of reconstructing the database in binary format isshown in the following Figure 3.

PROCEDURE ReconstructDB (database D, Unique G)

INPUT: Database D, Unique items G **OUTPUT:** Reconstructed DB in binary format**BEGIN:**

1. Load the input database D into the procedure
2. Load the Unique G into the procedure
3. Initialize Binary[ ] =
4. For each row ri R
5. For each Item Ii G
6. if [Ii= ri] do begin
7. Mark Ii "1" and store in Binary
8. Else
9. Mark Ii "0" and store in Binary
10. Close IF
11. Close For
12. Close For

Return Binary[ ]
END

Figure 3: Process of Reconstructing a Pseudo Code DB in Binary Format

Figure 3 shows the process of reconstructing a database in binary format using the following properties, "If the row identifier contains an item Xi then that item Xi = 1 else Xi = 0" The following table 3 shows the reconstructed database and this database is a and contains only zeros.

ITEM	T1	T2	T3	T4	T5	COUNT
M	1	1	1	0	1	4
N	1	0	1	1	1	4
O	0	1	1	1	1	4
P	1	0	0	0	0	1
Q	1	0	1	1	1	4

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

M	1	1	1	0	1
N	1	0	1	1	1
O	0	1	1	1	1
P	1	0	0	0	0
MNOP	0	0	0	0	0

Table 3: Reconstructed database

VII. SORT THE UPPER LEVEL ITEMS

The permutations start from the upper level (ie) {M, N, O, P, Q}, and the frequency or number of theseitems is calculated from Table 3, reconstructed using simple and function. For these binary values, as shown in the following section,
The items {{M, N, O, P, Q} are rare items with 0 support. Now the next level items are sorted and thecalculations are shown in the following section,

M	1	1	1	0	1
N	1	0	1	1	1
O	0	1	1	1	1
Q	1	0	1	1	1
MNOQ	0	0	1	0	1

The 4-item {M, N, O, P}, {M, N, O, Q}, {M, N, P, Q}, {M, O, P , Q}, {N, O, P, Q} and the count for these items is calculated using binary AND.

B	1	1	1	0	1	
N	1	0	1	1	1	
O	0	1	1	1	1	
P	1	0	0	0	0	
Q	1	0	1	1	1	AND
MNOPQ	0	0	0	0	0	

The set of items {M, N, O, P} is a set of rare items with 0 support. The next item is calculated.
Itemset {M, N, O, Q} =2 rare items with support. It is a non-zero rare item. The next set of items iscalculated.

The itemset {M, N, P, Q} = rare items with support 1 and is non-zero rare items.

M	1	1	1	0	1
N	1	0	1	1	1
P	1	0	0	0	0
Q	1	0	1	1	1
MNPQ	1	0	0	0	0

M	1	1	1	0	1
O	0	1	1	1	1
P	1	0	0	0	0
Q	1	0	1	1	1
MOPQ	0	0	0	0	0

M	1	1	1	0	1
N	1	0	1	1	1
O	0	1	1	1	1
MNO	0	0	1	0	1

M	1	1	1	0	1
N	1	0	1	1	1
P	1	0	0	0	0
MNO	1	0	0	0	0

The itemset {M, O, P, Q} =0 rare items with support and is zero rare items.
The itemset {N, O, P, Q} =0 rare items with support and is zero rare items. Now the next level 3-itempermutation is calculated as shown in this section.
3-Items {M, N, O}, {M, N,P}, {M, N, Q}, {M,O, P}, {M, O, Q}, {M,P, Q} , {N,O,P},{N,O,Q}, {N,P,Q},{O,P,Q}.

First {M, N, O} support is calculated as shown below,

N	1	0	1	1	1
O	0	1	1	1	1
P	1	0	0	0	0
Q	1	0	1	1	1
NOPQ	0	0	0	0	0

Items {M, N, O} are rare items with support =2. It is a non-zero rare item. The next set of items iscalculated.
Itemset {M, N, P} = rare items with support 1. It is a non-zero rare item. Similarly all permutations areobtained and calculated. The final results are shown in Table 4.

ITEMS	COUNT	ITEMS	COUNT
MNOPQ	0	MPQ	1
MNOP	0	NOP	0
MNOQ	2	NPQ	1
MNPQ	1	OPQ	0
MOPQ	0	MP	1
NOPQ	0	NP	1
MNO	2	OP	0
MNP	1	PQ	1
MOP	0	P	1
MOQ	2		

Table 4: Final output of rare itemset

VIII. EXPERIMENTAL RESULTS

The entire experiment is executed on a dual core processor with 4GB RAM at 2GHz and the datasets are collected from the following domain http://fimi.cs.helsinki.fi/data /. The proposed algorithm is developed using Java SMPF tool and compared with other algorithms like Apriori-Inverse [4] and ARIMA[5] and the results regarding runtime and memory consumption are shown in the following graphs. A synthetic database is created by the IBM-Quest tool and algorithms areexecuted on the synthetic database.

The characteristics of the ranked database are shown in the following Table 5, and these databases are used to implement the lgorithms.

Dataset	Size (MB)	# Items	# Trans	AvgTrans
Mushroom	19.3	119	8124	23
Retail	4.2	16,470	88,126	10.3
Pumsb	16.3	2,113	49046	74
Kosarak	30.5	41,271	990,002	8.1

Table 5: The characteristic of the database

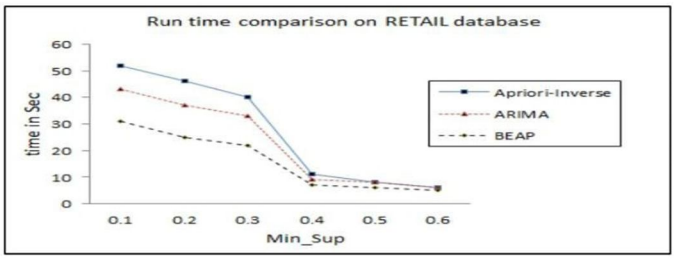


Figure 4: Comparison graph for run time on retail database

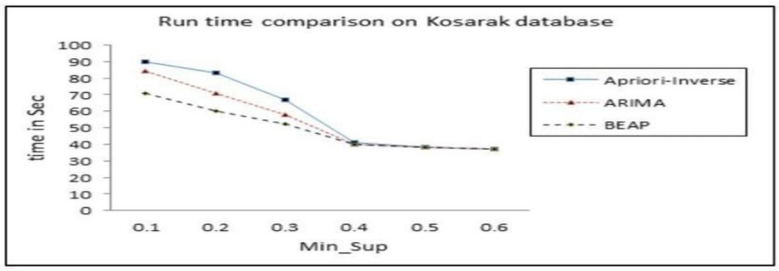


Figure 5: Comparison graph for run time on retail database

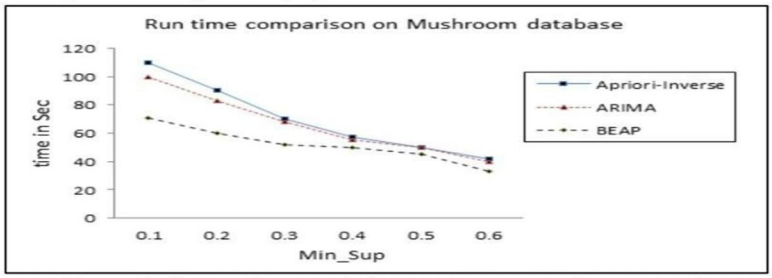


Figure 6: Comparison graph for run time on mushroom database

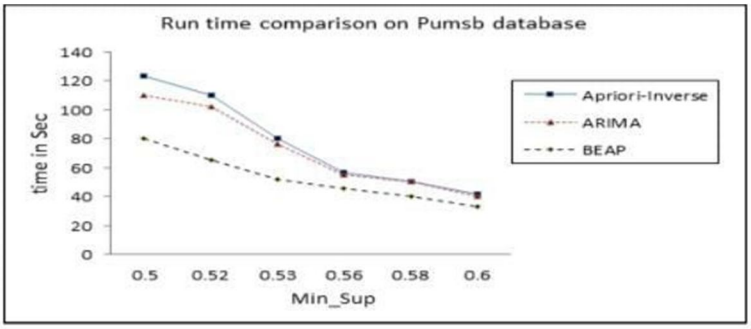


Figure 7: Comparison graph for run time on pumsb database

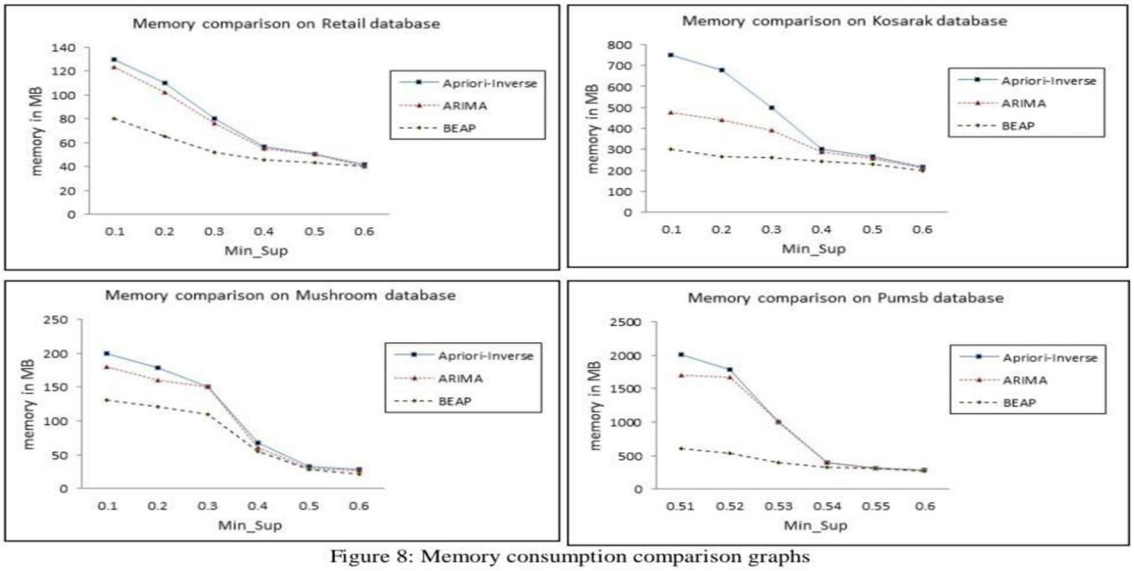


Figure 8: Memory consumption comparison graphs

From Figures 4, 5, 6, 7, it is clear that the proposed BEAP algorithm outperforms other existing algorithms in terms of running time by a good margin. The main reason for the performance of the proposed algorithm is that it only considered rare items and ignored other items, which significantly savedthe execution time and outperformed the other two algorithms. The next section shows the memory consumption.

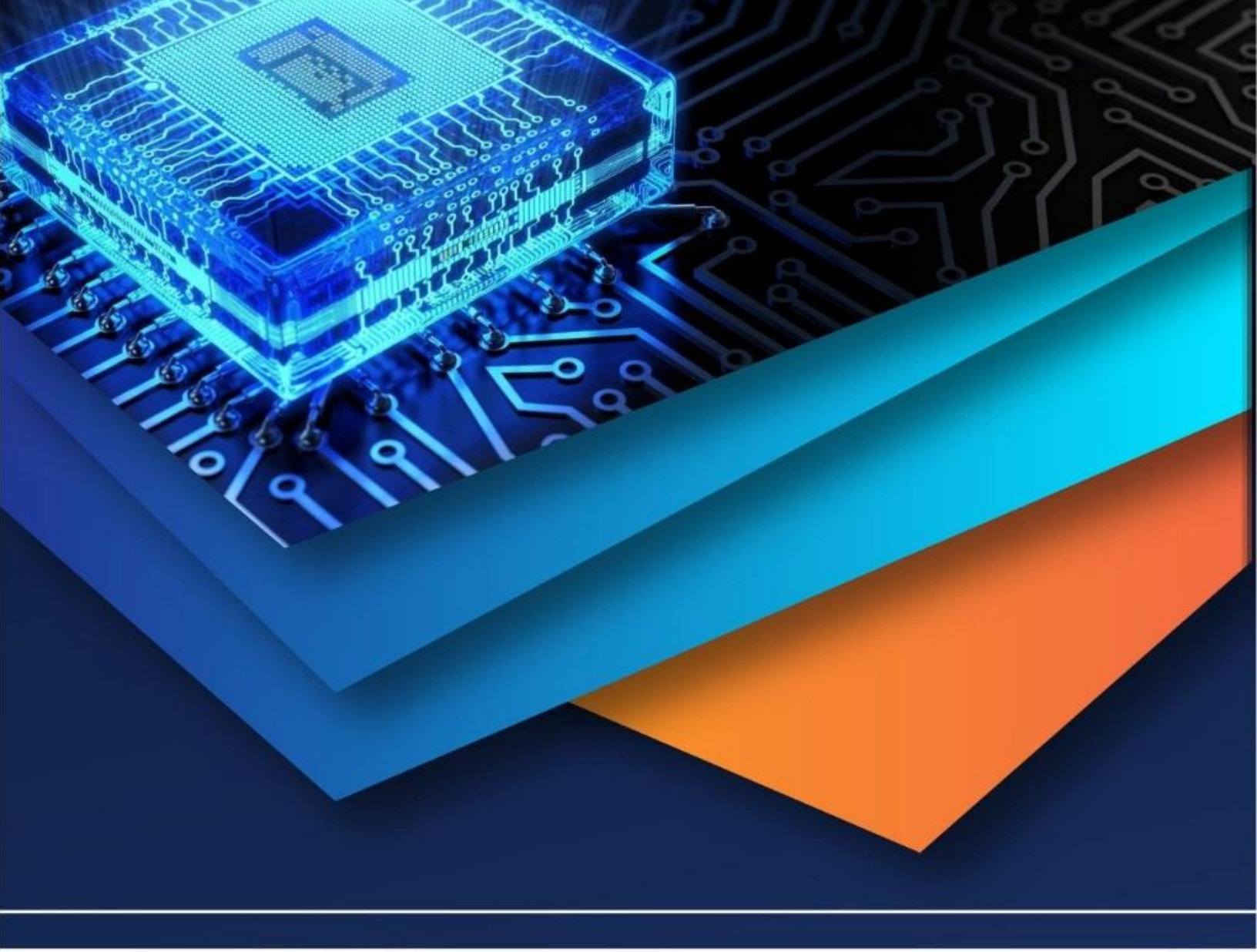
From Figure 8, it is clear that Aprioiri-inverse performed very poorly on a sparse database like Kosarak and min_sup is above 0.5, all algorithms performed equally well. When the algorithms are implemented on dense databases, the proposed BEAP algorithm outperforms the other two algorithms by a large margin. As shown in Figure 8, it is noted that the proposed BEAP algorithm consumed the least memoryfor all minimum support threshold values.

IX. CONCLUSION

In this paper a new algorithm BEAP is proposed to detect rare items and the proposed BEAP algorithm isa binary representation and below is an approach to detect rare items with minimal operation time and memory usage. The experimental results show that the proposed BEAP algorithm outperforms the other two algorithms in terms of speed and memory footprint.

REFERENCES

[1] E. Suzuki. An interesting exception is the unidirectional discovery of rules. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), 16(8):1065–1086, 2002.
[2] Wenke Lee and Salvatore J. Stolfo, “Data Mining Approaches to Intrusion Detection”, Security. , San Antonio, Texas, 1998.
[3] G. Weiss. Rarefaction Mining: An Integrated Framework. Explore SIGKDD. Newsl., 6(1):7– 19, 2004.
[4] Y. Koh and N. Rowntree. Finding Sparse Rules Using Apriori-Inverse. In Proc PAKDD '05,Hanoi, Vietnam, volume 3518 of LNCS, pages 97–106. Springer, May 2005.
[5] Laszlo Szathmary, Amedeo Napoli, and PetkoValtchev, “Towards Rare Itemset Mining,” 19th IEEE International Conference on Tools with Artificial Intelligence, Patras, Greece. 1, 2007, p. 305-312



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY