Performance Analysis of Rare Itemset Mining Algorithms

[1]Varsur Jalpa A., [2]Desai Sonali P., [3]Hathi Karishma B.
PG Student:CSE Department
B.H.Gardi College of Engg&Tech
Rajkot, Gujarat India

Abstract : Rare Or Infrequent itemset mining is a very important mining technique with wide applications. It found very useful in various domains like medical,biology,banking,retail,telecommunication industry,market basket analysis,predicting pre-birth terms etc.It discovers the hidden interesting association between data items.Infrequent itemset mining is a variant of frequent itemset also called rare itemset.Basic technique to find the correlation of data is Association Rule Mining.It has mainly two concepts Support and Confidence.Threshold plays a key role in deciding the interesting itemsets.The rare itemset may not found if high threshold is set.Some uninteresting items are appear if low threshold is set.for eg, Rare symptom which leads to disease. This paper focus on various mining algorithms like Apriori,FP-Growth,IWI Mining,MIWI Miner etc.

Keywords – Data Mining,Frequent Items,Infrequent Items,Threshold,Support,Confidence.

INTRODUCTION

Itemset is a set of items. Frequent items are appear very frequently in database,with high support and high confidence. Rare items are reverse from of frequent ones,it has low support and high confidence. [1]That has a vast range of real-life application. for eg,In **Medical**– if we identify the solution of rare disease then we can prevent the person to get affected by it.Here,we don't need this rare disease to become frequent.

In **Marketing** strategy knowing the rare items can help us to make them the frequent ones.So,the businessmen gains profit.

In **Market basket analysis** for finding which items tend to be purchased together,milk and bread occurs frequently and can be considered as regular case,some items like bad and pillow are infrequently associated itemsets.
Recently,the importance is being given for the discovery of infrequent or exceptional patterns[1].threshold plays key role for finding of frequent and infrequent itemsets.The things which are done together for eg.Buying groceries known as association,occurs between items.Association is an implication of the form X→Y. Infrequent itemsets are produced from very vast or enormous datasets.Extraction of frequent itemset is a necessary step in many association techniques[2].
Association rule mining extracts interesting correlation between transactions.In many application some items are appear very frequently in the data,while others rarely appear.if minsup is set too high,those rules that involve both frequent and infrequent items.To find the rule that contain both frequent and rare minsup is set to be very low.This may cause combinatorial explosion for those frequent items will be associated with one another in all possible ways.This problem is called the rare item problem[3].infrequent itemset do not comprise any infrequent subset.it appears only when threshold is set to very low.The mining algorithm solves the problem of discovering the infrequent itemsets in the given database[2].Infrequent Itemset Mining finds uninteresting items among the huge database.

A. Basic Concepts of Infrequent Itemset Mining[4]

Let $I = \{i_1, i2, \ldots, i_n\}$ be a set of data *items*, A transactional data set T={t1,t2,...,tn}is a set of transactions.and it is characterized by tid.An itemset I is a set of data items.we denote as **k-itemset**.
weighted transaction dataset Let I={i1,i2,..,im}.
Be a set of items. A weighted transaction dataset T is a set of weighted transactions.where ik,wk is weighted items.In general weights could be either positive,negative numbers or null.Itemset mined from weighted transaction dataset is known as weighted itemset.The problem of mining itemsets by considering weights associated with each item is known as weighted itemset mining problem.

Transaction T and IWI value of items are same then it is known as **equivalence property** or transaction equivalence.

The Rare Itemset Problem

In real-world databases are mostly non-uniform in nature containing both frequent and rare items.if the items frequency in a database vary widely,includes the two following issues:

i.if minsup is set too high,we will get the frequent items.
ii.if minsup is set too low this may cause combinational explosion,to produce rare items.This problem is known as rare item problem.[10]
The traditional algorithms determine valid rules by exploiting support and confidence requirements, and use a minimum support threshold to prune its combinatorial search space.[14]

METHODS OF INFREQUENT ITEMSET MINING

Rare Itemset mining is broadly classified into
Two groups:
a) Apriori Based(with candidate generation)
b)Pattern Growth Based(without candidate generation)

Apriori is well known algorithm.it derives some new algorithm.which are useful to find the frequent itemsets and rare itemsets.it uses the data structure like FP-tree and associate with items.it works with candidate generation.The mining of association rules for finding the relationship between data items in large databases is a well-known technique in data mining field with representative methods like Apriori [6]**.** ARM process can be decomposed as finding all frequent itemsets.

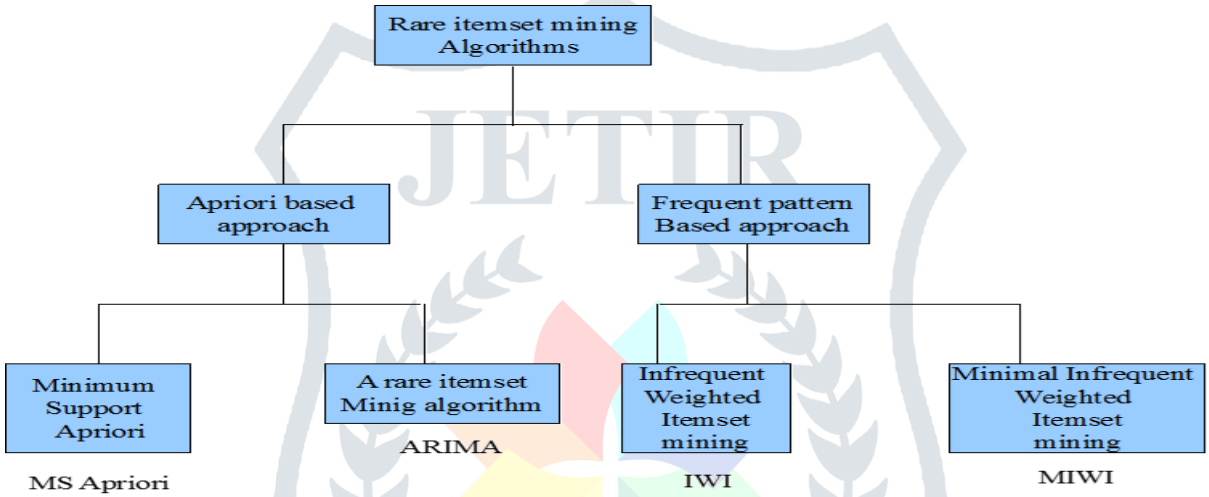

Fig.1 **Classification of Rare Itemset Mining Algorithm**

a) Apriori Algorithm

Apriori means a prior knowledge of data. It was the first proposed algorithm in association rule mining.it is based on iterative level wise search for finding frequent itemset generation[1].In that k specifies the largest set of items,and makes k passes over data.

It works with Generate and test Approach :for generate the candidate items and test if they are frequent.it is costly.by applying this technique we can also find the infrequent item.which are seems very less.large no. of candidate set is a main disadvantage of apriori algorithm.

Apriori Based Algorithms

Apriori cannot generate items with high confidence and this problem can be solved by specifying each item in dataset with a mininmum support given by user it known as MS Apriori(Multiple Support Apriori).[1]It is an extension of apriori algorithm.it improves the performance of apriori algorithm.
To generate rare items without need to specify the support threshold set by user,there is one another algorithm **RSAA**(Relative Support Apriori Algorithm).

In 2005, koh and rountree proposed another algorithm to find the rules that contain item over the maximum support threshold.that is **Apriori-Inverse algorithm.**

Another is **Apriori Rare Algorithm**.It has two main steps finding minimum rare item MIR,maximal frequent item MFI.it also find the generator of itemsets.by calculating threshold and scanning the database minimal rare items are generated.It is a modification of the Apriori algorithm used to mine frequent itemsets. Apriori-rare generates a set of all minimal rare generators, also called MRM, that correspond to the itemsets usually pruned by the Apriori algorithm when seeking for frequent itemsets. To retrieve all rare itemsets from minimal rare itemset (mRIs), a prototype algorithm called **"A Rare Itemset Miner Algorithm(Arima)**" was proposed.

Arima generates the set of all rare itemsets, splits into two sets: the set of rare itemsets having a zero support and the set of rare itemsets with non-zero support. If an itemset is rare then any extension of that itemset will result a rare itemset.[6]

The Apriori algorithms set the basic for a set of algorithms that relies largely on the apriori property and use the apriori- generate joint method to generate the candidate items. As per the apriori statement property all the nonempty subset from the frequent item set much also be the frequent.[11]

A Rare Itemset Miner Algorithm-it generates minimal rare items.it takes the mRI and produce rare items.main advantage of ARIMA algorithm it can find the rare items without zero itemsets.it depends on two threshold i.e, minsup and maxsup.[15]

Important terms of the apriori -based algorithms are[3]

1) *Breadth-first search technique used*: Basically the apriori based algorithms are work on this technique.
2) *Multiple scan of the database*: This feature is very unwanted because it requires the lots of processing time and IO cost.

*i)**Frequent Pattern Growth***-drawback of apriori algorithm is solved by frequent pattern growth.it works without candidate generation.Basically it works on divide and conquer strategy.it also called as fp-tree approach.it scan the database every times and collects the set of items.this algorithm reduce the no. of candidate generation and transactions.It works without candidate generation.It constructs fp-tree and calculate support of each item.then discard the frequent items.size of the tree will depends on the items.it is fast as compared to apriori.

Rare Association rule Generation-this method find the strong but rare associations as compared to local.the generated rules are known as mRI rules.

Positive and Negative Association Rules-it focused on identifying associations amonf frequent items.it find rules for both positive and negative items.it design the search space and increase the probability of items.

Residual tree-it is based on pattern growth approach,find minimaly infrequent items.it has no any subset.the residual tree concept is a variant of fp-tree.also called inverse fp-tree.in this method optimization of apriori algorithm is performed.[13]

b)Pattern-growth Mining Algorithms

The Pattern Growth algorithm comes in the early 2000s, for the answer to the problem of generates and test. The main idea is for to avoid the candidate generation step altogether, and to concentrate the search on a specified portion of the initial database. IWI (Infrequent weighted Itemset Mining):It is used to generate FP-tree associated with input weighted dataset T. FP-tree is initially occupy with set of equivalent transactions generated from T.it is a projection based approach.items belonging to header table is associated with the input.items are combined to generate new prefix.and the projected tree is recursively applied for mining infrequent itemsets. For example, by analysing clinical databases one can discover rare patterns that will help doctors to make decisions about the clinical care. In the security field, normal behaviour is very frequent, whereas abnormal or suspicious behaviour is less frequent.Considering a database where the behaviour of people in sensitive places such as airports are recorded, if we model those behaviours, it is likely to find that normal behaviours can be represented by frequent patterns and suspicious behaviours by rare patterns.[5]

MIWI(Minimal Infrequent Weighted Itemset Miner):It is similar to IWI Mining Algorithm.it mainly focus on generating only minimal infrequent items.The recursive extraction in MIWI Mining procedure is stopped as soon as infrequent items occur.

COMPARATIVE STUDY OF INFREQUENT ITEMSET MINING ALGORITHM

Comparative analysis of infrequent itemset mining algorithm is prepared on the basis of their different important features. For comparison infrequent itemset mining is subdivided into two broad categories, namely, Apriori Based and Pattern Growth Based Algorithms. All features used to categorize these algorithms are discussed first and then comparison is done for the following algorithms

IWI: Infrequent Weighted Itemset Mining
MIWI: Minimal Infrequent Itemset Miner
ARIM: A Rare Itemset Mining Algorithm
MS-Apriori: Multiple Support Apriori
RSAA: Relative Support Apriori Algorithm

Characteristics of Rare Itemset Mining Algorithm are:

Apriori vs. ECLAT

Apriori and ECLAT are the best known algorithm for mining frequent and rare items in a set of transactions.it achieves both memory and execution time.In that apriori uses fp-tree for demonstration and Eclat uses bit matrics for represent the itemset[9].

BFS-Based Approach Vs. DFS-Based Approach In a BFS approach level-by-level search can be directed to search the complete set of patterns i.e. All the children of a node are processed before proceeding to the next level. On the other hand, when using a depth-first search approach, all sub-arrangements on a path must be traversed before proceeding to the next one. The advantage of DFS over BFS is that DFS can very rapidly reach large frequent arrangements and therefore, some expansions in the other paths in the tree can be neglected.[13]

Top-Down Search Vs. Bottom-Up Search Apriori-based algorithms utilize a bottom-up search, enumerating every single frequent itemsets. This implies that in order to produce a frequent items will be produced. It can be easily concluded that this exponential complexity is limiting all the Apriori-based algorithms to find only short patterns, since they only implement subset infrequency pruning by eliminating any candidate sequence for which there exists a subsequence that does not belong to the set of frequent items.In a top-down approach the subsets of sequential patterns can be mined by constructing the corresponding set of projected databases and mining each recursively from top to bottom.[12]

Table1: Comparative study of itemset mining algorithms

Method	Input parameter	correctness	completeness	Candidate generation	Type of itemset	No.of DB scan	Approach
Apriori rare	minsup	✓	X	✓	Minimal rare,frequent	multiple	Bottom-up
ARIMA	minsup	✓	✓	✓	rare	multiple	Bottom-up
MS-Apriori	minsup	✓	✓	✓	rare	multiple	Bottom-up
Apriori Inverse	minsup	✓	X	✓	sporadic	multiple	Bottom-up
RSAA	minsup	✓	X	✓	rare	multiple	Bottom-up

Monotone Vs. Anti-Monotone Property In Anti-Monotone property states that every weighted transactional dataset less than a precedence relation holding between pairs of weighted itemsets.antimonotonicity is a property of support entails that if all subsets are frequent then itemset is also frequent[10].

Equivalence between IWI-support measure and Traditional support-measure It corresponds weighted transactional dataset that exclusively contain items I with weight w[4].

Bottom-Up,Divide and Conquer In this approach a global FP-Tree is generated from transaction database.it first divide the items then conquer into tree form.frequent items are generated adding items one-by-one.[16]

CONCLUSION

In this paper, we considered what is infrequent itemset mining and different types of their algorithms. This concept is being introduced in 1993[2] has gone through astonishing advancement in few years only. Initial work on this topic is focused on improvement of the performance of algorithms by using various data structure or various representations. So, on the basis of these problems the infrequent itemset mining is divided into two main groups, Apriori approach based algorithms and pattern growth approach based algorithms. From our comparative study and previous some studies by different research on itemset mining algorithms is found that the algorithm which are based on the approach of pattern growth are superior in terms of memory usage,speed and time-complexity.

REFERENCES

[1] K.Sadhasivam,Tamilarasi,"Mining Rare Itemsets with Automated Support Thresholds",*Journal of Computer Science*,pp.394-399,2011.

[2] Sonia Jadhav,G.M.Bhandari,"A Review on Efficient Mining Approach of Infrequent Weighted Itemset",*International Journal of Advanced Research in Computer Science and Management Studies*,vol.2,2014.

[3] Bing Liu,Wynne Hsu,"Mining Association Rules with Multiple Minimum Support",*International Conference on Knowledge Discovery & Data Mining*,1999.

[4] Luca Cagliero,Paolo Garza,"Infrequent weighted Itemset Mining Using Frequent Pattern Growth",*IEEE Transactions on Knowledge and Data Engineering*,vol.26,No. 4,April 2014.

[5] Mehdi Adda,Lei Wu,"Pattern Detection with Rare Itemset Mining",*International Journal On Soft Computing,Artificial Intelligence and Applications*,vol.1,No.1,August 2012.

[6] Jyothi Pillai,O.P.Vyas,"Overview of Itemset Utility Mining and Its Applications",*International Journal of Computer Application*,vol.5,No.11,August 2010.

[7] Nidhi Sethi,Pradeep Sharma,"Efficient Algorithm for Mining Rare Itemsets Over Time Variant Transactional Database",*International Journal Of Computer Science and Information Technologies*,vol.5,2014.
[8] Sujatha Kamepalli,Raja Rao,"Infrequent Weighted Itemset Mining in Complex Data Analysis",*International Journal Of Computer Applications*,vol.103-No.5,October 2014.
[9] Petko Valtchev,Amedeo Napoli,"Finding Minimal Rare Itemsets and Rare Association Rules",*International Conference on Knowledge, Science*,2010.
[10] Mehdi Adda,Lei Wu,"Rare Itemset Mining",*IEEE Sixth International Conference on Machine Learning and Applications*,2007.
[11] Lei Chen,He Jiang,"Incremental updating algorithm for infrequent itemsets on weighted condition",*International Conference on Computer Design and Applications*,2010.
[12] J.Jenifa,V.Sampath kumar,"Study On Predicting Various Mining Techniques Using Weighted Itemsets",*IOSR*,vol.9,pp.30-39,Mar-Apr 2014.
[13] B. Nath,N. Hoque,"A New Approach on Rare Association Rule Mining",*International Journal of Computer Application*,vol.53,September 2013.
[14] Ling Zhou,Stephen Yau,"Efficient association rule mining among both frequent and infrequent items",*Elsevier*,2007.
[15] Petko Valtchev,Amedeo Napoli,"Towards Rare Itemset Mining",*IEEE International Conference on Tools with Artificial Intelligence*,2007.
[16] Lugi Troiano,Cosimo Birtolo,"A Fast Algorithm for Mining Rare Itemsets",*IEEE International Conference on Intelligent Systems Design and Applications*,2009.

Author Profile

1)Varsur Jalpa is a student of Masters of Engineering in Computer Science and Engineering at B.H.Gardi College of Engineering and Technology, Rajkot, Gujarat, India. She is bachelors in Computer Science and Engineering. Her area of interest are Data Mining, Software Engineering and Information Security. Contact: +91 9925460983

2)Desai Sonali is a student of Masters of Engineering in Computer Science and Engineering at B.H.Gardi College of Engineering and Technology, Rajkot, Gujarat, India. She is bachelors in Computer Science and Engineering. Her area of interest are Data Mining, Computer networking and Security.Contact:+91 9408966536

3)Hathi Karishma is a student of Masters of Engineering in Computer Science and Engineering at B.H.Gardi College of Engineering and Technology, Rajkot, Gujarat, India. She is bachelors in Computer Science and Engineering. Her area of interest are Data Mining, Computer networking and Security. Contact:+91 9429810304