Indonesian Journal of Electrical Engineering and Computer Science
Vol. 13, No. 1, January 2019, pp. 41~47
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v13.i1.pp41-47

Design and development of an algorithm for mining rare itemsets

Sachin Sharma, Shaveta Bhatia
Faculty of Computer Applications, Manav Rachna International Institute of Research and Studies, Faridabad, India

Article Info

Article history:

Received May 16, 2018
Revised Jul 17, 2018
Accepted Jul 31, 2018

Keywords:

Apriori
Multiple minimum support threshold
Rare item set

ABSTRACT

Frequent item set mining emphases on excavating item sets which follow frequently in a transactional database. Investigators have projected numerous processes to mine rare item sets, however it is still stimulating. In current studies, to catch rare item sets, customer demarcated single threshold value ought be fixed small adequate which outcomes in generation of enormous quantity of duplicate item sets. The rare item sets may not initiate if a high threshold value is fixed. In this study, an exertion is generated to analyse the rare items for discovery of all probable rare item sets from the transaction database. To mine rare item sets effectually, a method has been anticipated that would permit customer to identify many lowest support that imitates the items and their diverse occurrences in list.

Corresponding Author:

Sachin Sharma,
Faculty of Computer Applications,
Manav Rachna International Institute of Research and Studies,
Faridabad, India.
Email: sachin.fca@mriu.edu.in

1. INTRODUCTION

Excavating patterns in professional relations, time-series, heritable archives, and several other classes of data is a simple stage in articulating propositions and realizing links among items. For example, the market basket exploration in the retail business is directed at determining which products is to be boughttogether in order to capture the buyinghabits of consumers or clients, to recognise their requirements, to improve cross-promotional series, to get new consumers or, additional in general, to develop corporate routines. In order to do this, usually experts look at the most persistent patterns, while also observing at the minimum repeated co-occurrences can deliver stimulating insights, regarding lesser but not less motivating groups[1].

1.1. Background

For generating rare itemsets, various researchers have designed the algorithms. Rare item problem is one of the significant challenges in association analysis. Hypothetically reviewing the rare item problem in association analysis started by Liu, Hsu, and Ma [2]. They anticipated a multiple minimum-support methodology in which every item in the data set is having its own minimum item support (MIS). MIS is computed by associating a smallest permissible support and the support of the item times a parameter, ß. In this means, rare items have smaller minimum-support associated to common items, therefore they will not be unnoticed in the rule generation process. There are two key hitches of this methodology. First, stipulating MIS when the amount of items in the dataset is huge is a monotonous work, and second, defining the optimum value of ß is not informal. To report these concerns, Szathmary and Valtchev [3] primarily uses Apriori-Rare which finds the Minimal Rare Item sets. ARIMA proceeds these Minimal Rare Items and gives the output as rare item sets. The key benefit is that the existing approach is able to catch rare item set

deprived of making zero item sets. However, it is based on two threshold standards, i.e. minimum support and maximum support.

A transaction mapping algorithm was designed by Song and Sanguthevar[4]. The main benefit of this algorithm is that it can generate rare item sets. Still, this algorithm is reliant on algorithms MRGExp and Apriori-Rare. In the paper [5], Arnab das tinted the significance of rare item sets in some areas. He anticipated an approach for determining the non-zero rare item set and generated motivating arrangements from various item sets.Yun, Ha, Hwang, and Ryu [6] presented relative support. Their procedure does not contain the factor ß, therefore they did not have the task of determining the optimum value for ß.

Geevlin and Mala [7-8] projected an algorithm for mining frequent item set and producing rare items from various data bases. In this study, two core glitches with current methods were projected. Excavating progression is being prepared with statistical methods like Poisson distribution.

Kiran and Re [9] anticipated a better multiple minimum-support method for mining the rare association rules. Their method needs specifying multiple minimum support, which is problematic related to a single minimum adjusted support. Srikant and Agrawalin 1997 produced a small alteration of classical algorithm. In this study, the rare items are moreover composed on the basis of minimum support value. Still, it flops to catch all the rare item sets.

Several investigators have applied association rules for diabetic patients. Piri et al [10] analysed the data of 23, 17, 259 patients identified with hypertension and diabetes. By put on association analysis, they initiate hypertension and diabetes was toughly connected. Mustafa et al [11] explored various algorithms for generating association rules. They analysed the algorithms on benchmark dense datasets and found FP-Growth algorithm is better as compared to others. Prasad [12] designed a new algorithm whose aim was to eliminate low-profit itemsets. The algorithm uses short time for generating high-profit itemsets. But the limitation is that it is unable to generate rare itemsets and may be extended for big datasets. Hussain et al [13] used Apriori algorithm for generating association rules and analysed the academic performance of students using WEKA.

To increase the quality of drawing out rare items, a technique named “Multiple Minimum Support Model” was developed by Darrab et al [14]. Every item is allocated along with a least support cost called “Minimum Item Support” (MIS) in this method. This value is allocated to every item identical to a section of its support. The approach, describes the least support of a rare rule in relations of minimum item support of the items that look in the rule. i.e. every item in the data base can have a least item support that can be considered by means of some method or can be quantified by the user. By giving various MIS values for various items, the user positively states different rules. This algorithm used the downward closure property. According to this property, every subset of frequent itemset is frequent. If we use this property, various interesting items may be ignored or discarded.

Alternative methodology for producing rare pattern are relative support Apriori algorithm (RSAA) proposed by Elahe et al [15]. This algorithm uses three customer stated supports known as First support, Second support and Relative support. If support value of any item is superior than or equal to first support value, it is called frequent item. If support value of an item is smaller than first support value but larger than or equal to second support value, it is called rare item; item sets having rare items have to mollify 2nd support and its relative support should fulfil lowest relative support identified by the user.

Bhatt et al [16] proposed an extension of RP-Tree algorithm called as Maximum Constraint Rare Pattern Tree Algorithm. This algorithm takes the transactional data set. With a previous MIS Value of item, this technique controls the rare item set from the data set. This tree chooses the transactions of single rare item set in it. The methodology finds only rare items and cuts the other item set from the transaction at the time of tree construction. This tree is the extension of RP-Tree While mining the rare itemsets, this algorithm uses tree generation. As an insertion of a node in the tree generation, the process may be expensive.

1.2. The Problem

In wholesale production [17], the market-basket study is wished for determining which things are to be bought together so as to keep buying behaviour of consumers. In market-basket study, some collections of things, such as toothpaste and toothbrush, happen commonly. When associated to milk & bread, some things like a chain & a gold ring are rarely related item sets, but reflected to be a significant relationship. We may also discover some infrequent relations that we cannot visualise. The problem of determining rare items has just caught the attention of the data mining.

Single minimum constraint model adopts that all items have analogous occurrence in the data set. In several existent applications, we will encounter following glitches [18]:

a) If the least support is fixed to a upper value, we are unable to catch the rules that contain of rare item sets.

b) For finding both rare and frequent items, it is necessary to fix little smallest support value but it makes a big quantity of common arrangements which are not valued.

Rare items can be used in several areas like text excavating - indirect relations can used to catch substitutes, antonym that is used in different situations. Infrequent patterns can be used to identify errors.

Infrequent item set has significant practice in [19]:

(i) Removal of undesirable association rules from rare item sets (ii) numerical disclosure risk valuation where rare patterns in unidentified survey data can lead to statistical revelation (iii) scam discovery where rare patterns in monetary data may propose uncommon activity related with fake behaviour (iv)Bio-informatics where rare patterns in microarray data may propose genomic disarrays. Rare items carry extremely stimulating information to several spheres with medication or natural science.

1.3. Proposed Solution

In this paper, the author will design a new algorithm/ technique whose purpose is to mine rare itemsets from the transactional database. This method will also overcome limitations of existing approach.

a) Apriori-Rare, Apriori-Inverse is not able to find rare item sets.
b) ARIMA Algorithm is able to generate rare item set but the rules produced from them are not all stimulating.
c) The process may be expensive as the existing algorithm uses tree generation method.
d) Various Rare itemsets may be ignored or discarded by using downward closure property.

The organization of the research paper is as follow: In section 2, proposed algorithm is offered and Investigational results are shown in section 3. Conclusion part is given in section 4.

2. PROPOSED METHOD

After reviewing and analysing the various algorithms, it is perceived that algorithms mentioned above have some drawbacks. The existing algorithms Apriori-rare, Apriori-inverse are not able to find all rare item sets. While the existing method/procedure ARIMA is able to generate rare item set. Also, the rules produced from them are not all stimulating. To remove the drawbacks met by the all methods, we suggest a new approach which follows a bidirectional approach. The new method uses the dataset and minimum threshold as input and yields rare item sets as output. It helps in pruning the candidate. The steps of proposed approach are as follow:

2.1. Steps:

a) Scan the database only one time to catchreal support of items.
b) Items in the transaction are in ascending/descending order according to multiple support thresholds.
c) Calculate MIS value for each item in transaction data set.

MIS = ß S(ij) if ß $S(i_j)$ > LS
$S(i_j)$ else

Where ß is a user defined value lies between zero and one;S (ij) denotes thepercentage support of any item equal to f (ij) / N*100; and LS is user defined least support value.

d) Find the least minimum support threshold.
e) Find rare item set if support is less than LS and larger than or equal to least minimum support threshold.
f) Find the transaction having atleast one rare item set from transaction data set.

2.2. Flow Chart

Following is the description provided for Given a transaction database DB as shown in Figure 1.

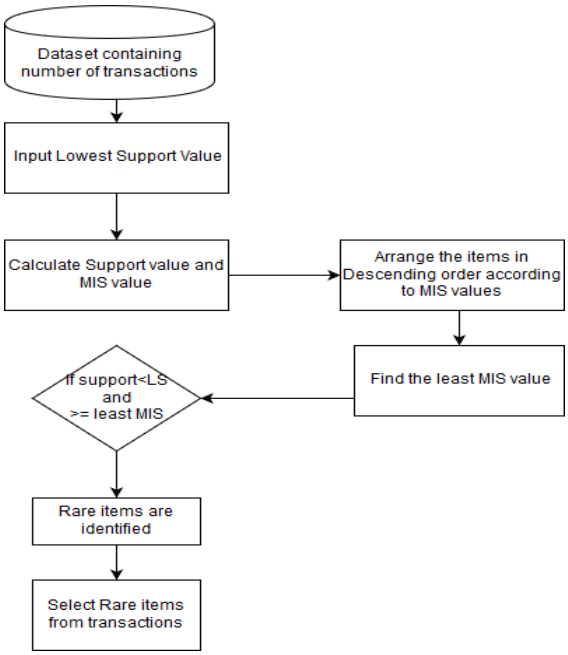

Figure 1. Given a transaction database DB as shown

Example: Given a transaction database DB as shown in the Table 1 with Minimum Support (LS) fixed at 20% and ß=0.7; the multiple item supportsof items in Table 2.

Table 1. Transactions Containing Various Items

TID	Items
T1	D,C,A,F
T2	G,C,A,F,E
T3	B,A,C,F,H
T4	G,B,F
T5	B,C

Table 2. Calculation of Support %age

Items	A	B	C	D	E	F	G	H
MIS	42	42	56	20	20	56	28	20
Support %age	60	60	80	20	20	80	40	20

As per the proposed approach, the actual supportpercentage and minimum support threshold values of various items calculated are shown in Table 1. MIS values are calculated according to the formula described in step 2 of proposed approach. Notice, if ß=1 and Sij ≥ LS, minimum support threshold value of any item are the real support of items Sij,whereas if ß=0, there is only single minimum support. It is to be noted that ß parameter is determined by the formula[9]:ß = $1/\acute{\alpha}$. In Table 2, the results are arranged and stored according to Minimum Support Threshold values.

Least minimum support threshold value is 20. D, E, H are rare item sets as these items are not smaller than least support value but larger than or equal to smallest minimum threshold value. So, in the next step, we will select the transactions having at least single rare item set. As shown in Table 4. In Table 3, the result is arrangement of items according of items according to minimum support threshold:

Table 3. Arrangement of Items According of Items According to Minimum Support Threshold

--	C	F	A	B	G	D	E	H
MIS	56	56	42	42	28	20	20	20
Support %age	80	80	60	60	40	20	20	20

Table 4. Selection of Rare Items from Transaction

Items	T1	T2	T3	T4	T5
MIS	D	E	H	-	-

3. RESULTS AND DISCUSSION

Here, the proposed method isequated with the existing versions, to discover rare item sets undermultiplesupportthresholds. To verify the efficacy and proficiency of the new method, several experiments are conducted using various data sets with different features. In these tests, we measure the performance with respect to time and memory space.

3.1. Experimental Environment and Datasets

We conduct the experiments by means of different kind ofdatasets to find the performance and efficiency of the proposed method. The data sets are executed and tested on machine Intel Core 2, 2.00 Ghz with 64 bit Operating system and are implemented in Python programming language.

We used three realworld datasets (Monk1, Chess and Cancer). The real world datasets are taken from FIMI repository [14]. The important characteristics real world datasets are given in Table 5-8

Table 5. Characteristics of Datasets

Data Set	Instances	Attributes
CHESS	3024	37
MONK1	432	5
CANCER	569	30

Table 6. Results Obtained for Chess

Data Sets	Time Execution	No of Rare item sets	Memory consumed (MiB)
200	0.85	25	1
1000	0.9	18	0.8
1500	0.87	17	0.8
3000	0.9	7	0.9

Table 7. Results Obtained for Cancer

Data Sets	Time Execution	No of Rare item sets	Memory consumed (MiB)
560	0.71	3	1

Table 8. Results Obtained for Monk1

Data Sets	Time Execution	No of Rare item sets	Memory consumed (MiB)
432	0.6	4	1

The execution time evaluation of proposed and various existing versions are given in Tables 9-10. The performance of proposed algorithm with severalß isdignified on the specified datasets. Note that, the execution time means the total runtime, which is the period among input and output. The experimental results divulge that proposed method is substantively faster than earlier versions.

Table 9. Comparison Between Proposed & Existing System (Chess Dataset)

	Time Execution(seconds)	No of Rare item sets
MSApriori	34	7
Existing [Gandhi P]	13	7
Proposed	0.9	17

Following is the description provided for comparison for Chess dataset as shown in Figure 2. Following is the description provided for comparison for Cancer dataset as soon in Figure 3

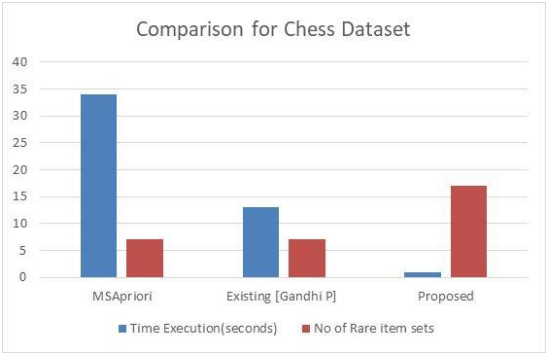

Figure 2. Comparison for chess dataset

Table 10. Comparison Between Proposed & Existing System (Cancer Dataset)

	Time Execution(seconds)	No of Rare item sets
Apriori	1	0
Existing [Hoque N]	1	0
Proposed	0.7	3

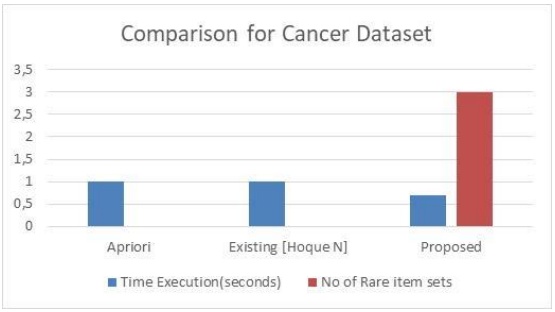

Figure 3. Comparison for cancer dataset

Comparison between proposed & existing system (monk1 dataset) as shown in Table 11.

Table 11. Comparison between proposed & existing system (monk1 dataset)

	Time Execution(seconds)	No of Rare item sets
Apriori	1	0
Existing [Gandhi P]	1.5	5
Proposed	0.6	4

Following is the description provided for Comparison for Monk1 data set as shown in Figure 4.

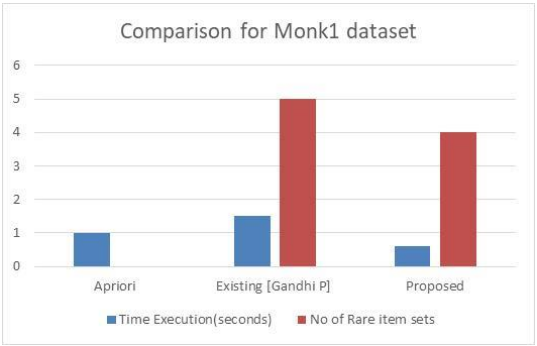

Figure 4. Comparison for Monk1 dataset

From the experimental results, it is observed that proposed algorithm gives better result as compared to previous algorithms. For Chess data set, it is found that proposed system gives better result i.e. number of rare items are high and time consumption is less. In case of Cancer data set, previous algorithms gives nothing but our method finds rare itemsets. However, for Monk1 dataset, the previous algorithm gives more rare itemsets than ours but it needs much amount of time and same number of database scan like Apriori.

4. CONCLUSION

As single minimum support is inadequate for association rule mining, it is unable to imitate frequency differences of the different items in the database. In realistic applications, such type of differences can be very huge. It is neither acceptable to set the minimum support too large, nor it is appropriate to set it too small. In this paper, we have explored the problem of using item specific minimum support. It permits the customer to stipulate multiple minimum item. To answer thisproblem, we have proposed an algorithm which is skilful of drawing out rare patterns efficiently. We have assessed the performance of proposed algorithm by showing atest on various datasets. The above mentioned results show that proposed algorithm has come out from the rare itemproblem and gives user more flexible and dominant model to state minimum support for rare item. Thus, proposed algorithm allows us to colliery rare pattern without creating any unexcitingand tedious pattern.

REFERENCES

[1] Troiano, Scibelli, Birtolo. "A fast algorithm for mining rare itemsets".*Proceedings of Ninth International Conference on Intelligent Systems Design and Applications*, 2009: 1149-55.

[2] Liu, Hsu, Ma, "Mining association rules with multiple minimum supports". *Proceedings of ACM SIGKDD International Conference on Knowledge discovery and data mining*, 1999: 337-341.

[3] Laszlo Szathmary, Amedeo Napoli, PetkoValtchev. "Towards Rare Itemset Mining".*19th IEEE International Conference on Tools with Artificial Intelligence*.2007. Patras: 305-312.

[4] Song, M., Sanguthevar, R. "A transaction mapping Algorithm for frequent item-sets mining".*IEEE Transactions on Knowledge and Data Engineering*, 2006;18(4): 472–481.

[5] Arnab Das. "Mining rare item sets using both Top-down and Bottom-up approach". *International Journal of computer science and information technologies.* 2016;7(3): 1607-1614.

[6] H. Yun, D. Ha, B. Hwang, K. H. Ryu. "Mining association rules on significant rare data using relative support". *Journal of Systems and Software-Elsevier*.2003; 67: 181-191.

[7] A.L. Greenie Geevlin, A. Mala. "Efficient Algorithms for Mining Closed Frequent Itemset and Generating Rare Association Rules from Uncertain Databases". *International Journal of scientific research and management*. 2013; 1(2): 94-108.

[8] Sethi, Sharma. "Efficient Algorithms for Mining Rare Itemset over Time Variant Transactional Database".*International Journal of computer science and information technologies*, 2014; 5(3): 3465-3468.

[9] R. UdayKiran, P. Reddy. "Mining rare association rules in the datasets with widely varying items' frequencies". *Proceedings of 15th international conference on database systems for Advanced Applications*. 2010: 1: 49-62

[10] Piri, Delen, Liu, Paiva. "Development of a new metric to identify rare patterns in association analysis: the case of analysing Diabetes com". *Expert systems with applications, Elsevier Ltd*, 2017;94: 112-125.

[11] Mustafa Man, Wan Bakar, Abdullah Z, Jalil M, Herawan T. "Mining Association rules: A case study on Benchmark dense data". *Indonesian Journal of Electrical Engineering and Computer Science*. 2016; 3(3): 546-553.

[12] K Rajendera Prasad. "Optimized High-Utility itemsets mining for effective association mining paper". *Indonesian Journal of Electrical and Computer Engineering*. 2017; 7(5): 2911-18.

[13] SadiqHussain, NeamaAbdulazizDahan, FadlMutaher Ba-Alwi, NajouaRibata. "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA". *Indonesian Journal of Electrical Engineering and Computer Science*. 2018; 9(2): 447-459.

[14] Darrab, Ergenc. "Vertical Pattern Mining algorithm for multiple support thresholds". *Proceedings of International conference on knowledge based and intelligent information and engineering Systems*. France. 2017: 417-426.

[15] Elahe, Zhang. "Mining frequent itemsets along with rare itemsets based on categorical multiple minimum support". *Journal of Computer Engineering*, 2016; 18(6): 109-114.

[16] Bhatt, Patel, "A Novel approach for finding rare items based on multiple minimum support framework". *Proceedings of 3rd International conference on recent trends in computing*. 2015; 57:1088-1095.

[17] KanimozhiSelvi. "Mining Rare item set with automated support thresholds". *Journal of Computer science*, 2011;7(3): 394-399.

[18] Srikant, R., Agrawal, R. "Mining Generalized Association Rules". *Future Generation Computer Systems*, 1997; 13: 161–180.

[19] Padmavathy, Joe. "Rare utility itemset mining without candidate generation". *International Journal of Advanced Research in Management, Architecture, Technology and Engineering*. 2016;2(12): 121-128.