IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 2, Ser. II (Mar. –Apr. 2022), PP 15-21
www.iosrjournals.org

Mining Multidimensional Rare Association Rules

SunithaVanamala[1], L. Padma sree[2], S. Durga Bhavani[3]
[1]Assistant Professor, Dept. of Computer Science, Lecturer, Department of CS, TSWRDCW, Warangal East, Telangana, India
2 Professor, Department of ECE, VNR Vignan Jyothi Institute of Technology and science, HYD
[3]Retd. Professor, School of Information Technology, JNTUH ,HYD

Abstract: *In data mining, association rule mining is a very popular and well-studied method to mine interesting relations among attributes in large databases. The evolution of various new application domains, such as bioinformatics and e-commerce, emphasis on analyzing high dimensional data has begun. Extracting the important relationships in data is a big challenge in the case of high-dimensional data. Interesting associations among dimensions based on frequent predicates are well studied, but research on associations among rare or infrequent dimensions is not much explored, there are applications like credit card fraud detection, network intrusion detection where rare associations among attributes trigger the system to discover fraudulent cases. In this paper, we are presenting an Eclat-based multi-dimensional rare association rule mining algorithm. As dimensions in the transactions may be categorical or numeric, the numeric attribute values are discretized using equal frequency binning. The proposed method is the first of its kind, has shown good performance.*

Keywords: *Association rules, rare itemsets, infrequent itemsets, multi dimensional association rules, Eclat, bitsets.*

Date of Submission: 10-03-2022 Date of Acceptance: 26-03-2022

I. Introduction

Knowledge discovery from large transactional databases using various data mining techniques is widely used for decision making in many applications such as sentiment analysis, trend analysis, recommendation systems, and fraud detection. Association rule mining techniques [1-6] find a relationship between itemsets in data, clustering, and classification are techniques to group the unlabeled data into clusters that identify the different classes existing in categorical data. Association rules mainly from frequent itemsets is widely used for market basket analysis, however, associations among rare itemsets and rare & frequent itemsets are also very useful in many applications like anomaly detection, fraud detection, etc. Rare associations emphasize unusual patterns rather than common or regular patterns.

In this paper, we propose a method to identify multidimensional rare association rules using Eclat based approach which is an extension of Eclat_RPGrowth[7]. The paper is organized as follows, Literature study is discussed in the next session, section 3 explains the proposed model, Experimental study and its analysis is covered in the section-4 and conclusion and future work discussed in section-5.

II. Related Work

Association rule mining is an emerging research area in data mining that is based on frequent or rare itemsets. Finding association or correlation among frequent or rare itemsets is useful in many different applications such as market basket analysis, recommender systems, intruder detection systems, etc.

Rare items are the items that occur with less frequency but the confidence of association with other items is very high. Finding such associations is an important data mining task, as it is used in many application areas like fraud detection, disease prediction based on rare symptoms association, network intruder detection based on log entries, etc.

Many authors have proposed different algorithms to mine rare patterns based on Apriori [1,3-6], FPtree[2] a few algorithms were also proposed that are based on Eclat [7-9] which is a vertical mining algorithm

Association among items in a single dimension is general association rule mining, but real time applications involve more than one dimension, so for decision making in some applications requires multi dimensional association. Dimensions/Attributes can be categorical or Numeric, Categorical attributes have a finite number of possible values, with no ordering among the values, such as color and gender. Quantitative attributes are numeric and have an implicit ordering among these values. Hence numeric attributes are to be transformed to nominal attributes to find associations among frequent or rare itemsets.

To generate patterns with Quantitative attributes, the attribute values must be preprocessed i.e each numeric attribute has to be discretized. In literature, there are different approaches in mining multi dimensional association rules

Mining using Static Discretization: Discretization is static and occurs prior to mining, Discretized attributes are treated as categorical

Mining using Dynamic Discretization, Known as Mining Quantitative Association Rules Numeric attributes are dynamically discretized, Rules involving more than one dimensions or predicates

- buys (P, "DELL Laptop Computer") -> buys (**P, "SONY Inkjet Printer"**) (Single dimensional)
- **age (P, "20 ..25") and occupation (P, "student") – > buys (P "Canon Inkjet Printer")** (Multi Dimensional- Inter dimension Association Rule)
- *age (P, "20 ..25") and buys (P, "SONY Laptop Computer") – > buys (P, "SONY Inkjet Printer")* (Multi dimensional- Hybrid dimension Association Rule)

A Boolean matrix [13]-based approach is employed to produce a multidimensional rule with no repetitive predicates. The items comprising a rule come from multiple dimensions, hence a Boolean Matrix-based technique was utilized to determine the frequent itemsets. It's a method for extracting multidimensional association rules from relational databases using an algorithm. To find common predicate sets, the technique uses Boolean relational calculus. When you initially use this algorithm, it will scan the database once and build association rules. The Apriori property is used in the item set pruning process. The association rules are generated using Boolean logical functions, thus there is no need to examine the database again. Because it saves all data in bits, it requires less memory.

The implementation of the Ali, Moniri [14] association rule algorithm is demonstrated using a netting plant's net producing process. This method resulted in a considerable reduction in the numeber of flaws in the manufactured nets.

In Pandey and Pardasani's [16] method, if the dimensions come from various tables, combine them. To begin, combine the participant tables into a general table in order to construct rules that express the relationship between two or more domains. The mapping code is then applied to the selected dimension, which may then be uploaded to the information system as a single attribute. In the second stage, candidate itemsets are constructed using equivalence classes and also translating the mapping code into real dimensions in order to determine the association rules.

The Sridevi and Ramaraj [19] method generate the data set for the transactional database based on the attribute. To categorize the attribute value according to the restrictions, Data Discretization and Concept Hierarchy are used. The quantitative conversion is used for numeric attributes

Cui et al. [15] presented FRI-Miner, which uses fuzzy theory with linguistic meaning to find valuable and intriguing fuzzy rare itemsets in a quantitative database. FRI-Miner also employs a fuzzy-list structure to hold important data and employs a number of pruning algorithms to decrease the search space.

As there are many proposed approaches to deal with multidimensional frequent itemset mining, but much focus is not given to find multi dimensional rare association rule generation, a method to mine multidimensional rare association rules from transactional data bases is proposed in this paper.

III. Proposed Work

A. Basic Concepts and Definitions:

The set of input transactions in the database represented as TDB = {T1,T2,T3,T4…..Tn}, Where T1,T2 are the transactions such that 1<=i<=n. Each transaction may have k items, 1<=k<=m, Item set I= {I1,I2,I3,…Im}, where m is the distinct items present in the database. Itemset: Ex. {X,Y} is a representation of the list of all items which form the association rule

1) *Support: An itemset X Є I, then support of X is the fraction of transactions in the database that contains itemset X.*
2) *Rare Item: An itemset X is rare item if Supp(X)<= MinFrequentSupportThreshold (MFT) and Supp(X)> MinRareSupportThreshold (MRT).*
3) *Frequent Item: An itemset X is frequent item If Supp(X)>= MinFrequentSupportThreshold (MFT)*
4) *Perfect Rare ItemSet: itemset X ЄI is called perfect rare ItemSet if all items in the X are rare items.*
5) *Rare item itemset: An itemset is called a rare item itemset if it has at least one rare item in it.*
6) *Association Rule: Ex. {X → Y} is a representation of finding Y on the basket which has X on it*

7) *Confidence: Probability of occurrence of {Y} given {X} is present, MCT(minimumCondidence threshod is used*
8) *Lift: Ratio of confidence to baseline probability of occurrence of {Y}*

B. Mining multidimensional Rare Itemsets using modified Eclat_RPGrowth:

In our previous paper Eclat-RPGrowth[7], proposed the algorithms using vertical data format with BitSet. Each bitset represents transaction ids with bits, the size of the bitset vector is equivalent to a number of transactions in the database, and each bit corresponds a transaction. The tree structure used consists of a set of levels, depending on the maximum transaction length or user specified maximum pattern length. Each level again is a collection of tree nodes. Each node is a 6 tuple-{itemset(IS), BitSet(BS), ParentNode1(P1), ParentNode2(P2), childNodesPtrList(CPtrs List), isVa-lid(IV).

The proposed method is an extension of Eclat_RPgrowth, the tree structure used is similar to the structure described, but an additional data member, attributeGroupidlist is added, it is used within add item method to skip unnecessary join operations among predicates related to same dimension and the proposed algorithm considers categorical and Numeric attributes. Categorical attributes are like gender with values male or female where the the attribute has more than one value and there is no significance for the ordering of values of an attribute. A numeric value of the attribute is like age and income are continuous values, order of values is also important. To perform pattern mining, numeric attributes cannot be directly used, as it gives so many possible values, hence we discretized the numeric attributes using equal frequency binning.

.

C. Steps in the algorithm:

The following procedural steps describe the process required to generate the multidimensional rare association rues from the database.

Input: Transaction file, Attribute Information file, gb (MRT ,MFT, MCT)

1. Preprocess Transaction database: numeric data has to be discretized using equal frequency binning.

3.Create globals object gb (MRT,MFT,MCT,)

4. Create an object of EclatRarePatternTree as erp and object of DatasetAttributesMap as dsAttributeMap

5. call method createItemSetIdMap

i) Assign dimension_id for each attribute in the database, used to avoid candidate generation among subordinate attribute-value pairs, adds it to temSetIdMap

ii) For each discretized/categorical value of attribute/dimension create a new binned item or predicate, assign id, eg male is given the value 1, female is given the value 2, as each dimension may have multiple predicates, and add it to corresponding itemNumNameMap, update item's groupid(dimension id) in the itemAttribGroupIdMap.

6. Find rare predicates that satisfy support criteria

i) call erp.runEclatRplusGrowth(filename, gb, dsAttributeMap)

ii)erp.oneItemList= createRPlusOneItemVMList(Sourcefile,gb);

iii) **for**(**int** i=0;i<oneItemsList.size();i++)

```
{
 sItem=oneItemsList.get(i);
erp.addItemToEclatRPlusTree(sItem, gb, dsAttribMap);
}
```

7. create object of RareItemsetData as data

8. call data.getRareItemsdatafromErp(erp);

9. Generate rare association among predicates with high confidence and positive lift value.

i)create AssociationRuleGeneneratotr object as associationRulesGen

ii) create ArrayList<AssociationRule> associationRuleList

iii)call associationRulesGen.mineAssociationRules(data, gb.MCT);

10. create object of SortedReadableRules as SRR

11 SRR.getSortedReadableRules(associationRuleList, dsAttributeMap);

12 SRR.displayRules();

Table No 1: Sample Data Base

Tid	Location	First language	Gender	Age	Annual Income	Result	Course
1.	Rural	Telugu	M	20	1,45,000	Pass	Arts
2.	Rural	Telugu	F	19	1,55,000	Pass	Arts
3.	Urban	English	M	21	93,000	Fail	Science
4.	Rural	Telugu	M	18	95,000	Pass	Science
5.	Urban	English	F	19	1,20,000	Fail	Arts
6.	Urban	English	F	21	1,60,000	Pass	Arts
7.	Rural	Telugu	M	22	94,000	Pass	Science
8.	Rural	Telugu	M	18	1,22,000	Pass	Science
9.	Urban	English	F	19	1,15,000	Fail	Arts

The result of the algorithm is explained with the sample database shown in Table 1, The parameters, MFT as 0.5,MRT as absolute support value 5 and MCT as 0.8.Data is discretized using WEKAa with equal frequency binning, where age is discretized to 3 bins and income into 3 bins.
Step1:@ATTRIBUTE=Age=ENUMERATION= (-inf-18.5]=(18.5-19.5]=(19.5-inf)
@ATTRIBUTE=Annual_Income=ENUMERATION= (-inf-95500]=(95500-133500]=(133500-inf)
Step 2: Atrributegroupid is assigned to each dimension.
Step3: Each attribute- value (predicate) pair is given an attribute id as shown in Table 2
Step 5: describes the how to avoid redundant predicates while generating large itemset.
Step 6: describes the process used to generate one items & large itemsets using vertical mining the process is same as described in reference [7]
Steps 8-9: describes the rare association rule generation

TableNo2: Mapping items to item-value pair predicates

@ITEM=1=Location= Rural @ITEM=2=Location= Urban @ITEM=3=First_language= Telugu @ITEM=4=First_language= Telugu @ITEM=5=Gender= F @ITEM=6=Gender= M @ITEM=7=Age= (-inf-18.5] @ITEM=8=Age= (18.5-19.5]	@ITEM=9=Age= (19.5-inf) @ITEM=10=Annual_Income= (-inf-95500] @ITEM=11=Annual_Income= (95500-133500] @ITEM=12=Annual_Income= (133500-inf) @ITEM=13=Result = Pass @ITEM=14=Result = Fail @ITEM=15=Course = Arts @ITEM=15=Course = Science

TableNo3: Transaction table with assigned iem ids

Tid	ITEMS	Tid	ITEMS
1	1,3,6,9,12,13,15	5	2,4,5,8,11,14,15
2	1,3,5,8,12,13,15	6	2,4,5,9,12,13,15
3	2,4,6,9,10,14,16	7	1,3,6,9,10,13,16
4	1,3,6,7,10,13,16	8	1,3,6,7,11,13,16

TableNo4: Support values of one itemsets

Item no	Support Count	Support	Fr/Rare	Item no	Support Count	Support	Fr/Rare
item: [1]	6	0.6	Frequent	item: [9]	4	0.4	Rare
item: [2]	4	0.4	Rare	item: [10]	3	0.3	Rare
item: [3]	6	0.6	Frequent	item: [11]	4	0.4	Rare
item: [4]	4	0.4	Rare	item: [12]	3	0.3	Rare
item: [5]	4	0.4	Rare	item: [13]	7	0.7	Frequent
item: [6]	6	0.6	Frequent	item: [14]	3	0.3	Rare
item: [7]	2	0.2	Rare	item: [15]	6	0.6	Frequent
item: [8]	4	0.4	Rare	item: [16]	4	0.4	Rare

The resulting transactions are shown in Table 3 . Total number of Multi dimensional rare association rules generated is 158. The sample rules given here describe the combination of predicates that caused the failure in semester exams.
[4, 5, 8] -> [2, 14]   cnf: 1.0   lift: 3.3333333333333335
[4, 8, 15] -> [5, 14]   cnf: 1.0   lift: 5.0

[4, 8, 15] -> [5, 14]   cnf: 1.0   lift: 5.0

II. Experimental Study And Discussion

The study was carried out on real world datasets 1. Dermatology data set from UCI machine learning repository [11], 2. Diabetes data set from WEKA datasets [12]. The experiments were conducted on an Intel Core i5 2.4 GHz machine running under the Windows 10 Operating system with 8 GB RAM. The value of MFT is assumed as 15%, 12%, and 20% and absolute support for MRT as 5. The number of rare itemsets generated are compared in the graph shown in figure I. The results of execution under different parameter settings are shown in Table 6&7. The analysis of results of the dermatology data set shows that 99.5%-99.9% of the generated rules are positive with an MCT value of 0.9(90%) and 98%-99% of the generated rules are positive with a MCT value 0.8(80%). The method generated the most significant rules with an MCT value of 0.9, 25%-40% additional rules are generated when the MCT value is reduced to 0.8(80%).

The analysis of the results of the diabetes data set shows that 100% of the generated rules are positive with an MCT value 0.9(90%) and 0.8(80%). The method generated the most significant rules with an MCT value of 0.9, the number of generated rules are almost doubled when the MCT value is reduced to 0.8(80%). When the rules are large in number, it is difficult to understand rules and relationship among attributes, depending on the application in the context the user can set the optimal value for MCT based on domain knowledge

TableNo5: Data Sets

Data set	Number of Instances	Number of Attributes
Dermatology	366	34
Diabetes	768	9

Table No6 : **Comparison of Result of Multidimensional Eclat_Rpgrowth W.R.T Different Parameters (MFT, Confidence, Lift) on Dermatology Data Set**

Sno	MFT	Confidence	Number of Rare itemsets	Positive rules	Negative rules	No correlation rules	Total No.of rules	Execution time (Seconds)
1	0.12	0.9	5761	3515	15	0	3530	1.135
2	0.12	0.8	5761	4690	116	0	4806	1.403
3	0.15	0.9	19654	16615	37	0	16652	3.863
4	0.15	0.8	19654	26729	382	0	27111	3.941
5	0.2	0.9	58930	88402	130	0	88532	18.584S
6	0.2	0.8	58930	137992	1126	0	139118	25.903

Table No 7: Comparison of result of Multidimensional Eclat_RPGrowth w.r.t different parameters (MFT, Confidence, Lift) on Diabetes data set

Sno	MFT	Confidence	Number of Rare itemsets	Positive rules	Negative rules	No correlation rules	Total No.of rules	Execution time (Seconds)
1	0.15	0.9	4015	353	0	0	353	0.666
2	0.15	0.8	4015	653	0	0	653	0.623
3	0.2	0.9	4138	407	0	0	407	0.59
4	0.2	0.8	4138	719	0	0	719	0.862
5	0.12	0.9	1065	31	0	0	31	0.293
6	0.12	0.8	1065	58	0	0	58	0.424

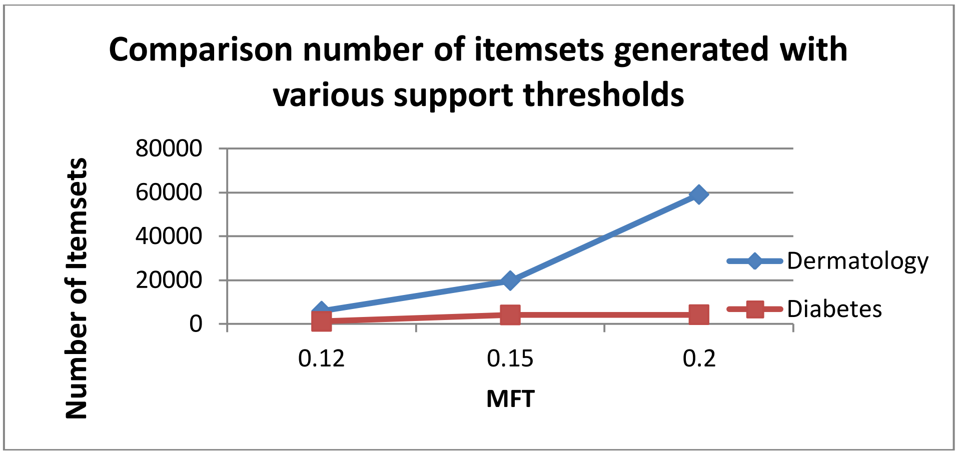

Fig. 1. Comparison of itemsets generated with various support thresholds

III. Conclusion And Future Scope

This study used vertical mining with bitsets and a breadth-first and depth first rare pattern tree to provide a new method for identifying multi dimensional rare itemsets from databases. The two user-defined support thresholds MFT and MRT were used in this study.MRT is used to discard the absolute noise itemsets. BitSet representation for vertical data format enhanced the performance of intersection operation. The mining least association rule is definitely important in determining rarity.

In the database, if there is an irregular relationship between the itemsets, It's a complicated situation, It's highly difficult and computationally expensive to identify rare associations among multiple attributes, and necessitates careful measurement to capture the fewest rules. To reduce the complexity, a method which is an improved version of Eclat_RPGrowth is proposed, specifically useful in detecting the irregular associations in the medical environment, can sometimes save a person's life. So to analyze the performance, experiments were carried out on Dermatology and diabetes data set. In addition to support thresholds, confidence and lift measures are used to capture more significant association rules. The outcome demonstrated improved performance in terms of execution speed and quality of generated rules. Other interesting metrics may be applied to further reduce the number of rules, in our future work we will use different measures to improve the performance.

References

[1]. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases.In: Proceedings of 20th international conference on very large data bases (VLDB). VLDB, pp 487–499

[2]. Tsang, S., Koh, Y.S., Dobbie, G.: RP-Tree: Rare Pattern Tree Mining. In: Cuzzocrea, A., Dayal, U. (eds.) DaWaK 2011. LNCS, vol. 6862, pp. 277–288. Springer, Heidelberg (2011)

[3]. SunithaVanamala, L.Padma Sree, S.Durga Bhavani ,"Efficient rare association rule minig algorithm " Int. J. of Engineering Research and Applications (IJERA) , Vol. 3, Issue 3, pp.753-757 , 2013.

[4]. Sunitha Vanamala., Sree, L., & Bhavani, S. (2014). Rare association rule mining for data stream. International Conference on Computing and Communication Technologies, 1-6.

[5]. Manal Almuammar, Maria Fasli, "Learning Patterns from Imbalanced Evolving Data Streams", Big Data (Big Data) 2018 IEEE International Conference on, pp. 2048- 2057, 2018.

[6]. Manal Almuammar, Maria Fasli, "Pattern discovery from dynamic data streams using frequent pattern mining with multi-support thresholds", the Frontiers and Advances in Data Science (FADS) 2017 International Conference on, pp. 35-40, 2017.

[7]. Sunitha Vanamala., Padma Sree L., Durga Bhavani S. (2021) Eclat_RPGrowth: Finding Rare Patterns Using Vertical Mining and Rare Pattern Tree. In: Pandian A., Fernando X., Islam S.M.S. (eds) Computer Networks, Big Data and IoT. Lecture Notes on Data Engineering and Communications Technologies, vol 66. Springer, Singapore. https://doi.org/10.1007/978-981-16-0965-7_14

[8]. Zaki M (2000) Scalable algorithms for association mining. IEEE Trans Knowl Data Eng 12(3):372–390

[9]. Ma Z, Yang J, Zhang T, Liu F (2016) An improved Eclat algorithm for mining association rules based on increased search strategy. Int J Database Theory Appl 9:251–266

[10]. Darrab S, Ergenic B (2017) Vertical pattern mining algorithm for multiple support thresholds.In: International conference on knowledge based and intelligent information and engineering (KES). Procedia computer science, vol 112, pp 417–426

[11]. Frank A, Asuncion A (2010) UCI machine learning repository. http://archive.ics.uci.edu/ml

[12]. Frequent itemset mining dataset repository. http://fimi.uantwerpen.be/data/

[13]. N. Khare, N. Adlakha and K. R. Pardasani, "An Algorithm for Mining Multidimensional Association Rules Using Boolean Matrix," *2010 International Conference on Recent Trends in Information, Telecommunication and Computing*, 2010, pp. 95-99, doi: 10.1109/ITC.2010.8.

[14]. Ali, Seyed, Alireza Moniri, and Farshad Mohebbi. 2017. "Root-Cause and Defect Analysis Based on a Fuzzy Data Mining Algorithm." International Journal of Advanced Computer Science and Applications 8(9). (February 12, 2022).

[15]. Cui, Yanling, Wensheng Gan, Hong Lin, and Weimin Zheng. 2021. "FRI-Miner: Fuzzy Rare Itemset Mining." arXiv:2103.06866 [cs]. http://arxiv.org/abs/2103.06866 (February 2, 2022).

[16]. Pandey, Anjana, and KamalRaj Pardasani. 2009. "Rough Set Model for Discovering Multidimensional Association Rules." .

[17]. Sathiyapriya, K et al. 2012. "Privacy Preserving Quantitative Association Rule Mining." :

[18]. Sharma, Rakesh, and Pinki Sharma. 2011. "Mining Multidimensional Association Rules." International Journal of Advanced Research in Computer Science 2(4): 100–103.
[19]. Sridevi, R., and E. Ramaraj. 2014. "Multidimensional Quantitative Rule Generation Algorithm for Transactional Database." International Journal of Computer Applications 99(2): 40–44.
[20]. Usman, Muhammad, Russel Pears, and A.C.M. Fong. 2013. "Discovering Diverse Association Rules from Multidimensional Schema." Expert Systems with Applications 40(15): 5975–96.

SunithaVanamala, et. al. "Mining Multidimensional Rare Association Rules." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 24(2), 2022, pp. 15-21.