International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-2, Issue-4, July-2014

A Perspective on Boolean Matrix for Mining Hybrid Dimensional Association Rules

Chaitanya B. Pednekar, Prof. R.C.Suryawanshi

***Abstract*— The discovery of association rules in data mining is an important issue, the core of which is the frequent pattern mining, Apriori algorithm is traditional for the association rule mining, but it should repetitively scan the database and can produce number of candidates. We present an algorithm of mining hybrid dimension association rules which satisfies the definite condition on the basis of multidimensional transaction database. Boolean Matrix based approach has been employed to generate frequent item sets in multidimensional transaction databases. When using this algorithm first time, it scans the database once and will produce the association rules. Apriori property is used in algorithm to prune the item sets. It is not needed to scan the database again; it uses Boolean logical operations to generate the association rules. It is going to store data in the form of bits, so it needs less memory space.**

Keywords—Hybrid dimensional association rule, Frequent itemsets, Boolean matrix, multidimensional transaction database.

I. INTRODUCTION

Finding frequent patterns plays an important role in data mining and knowledge discovery techniques. Association rule describes correlation between data items in large databases or datasets. The first and foremost algorithm to find frequent pattern was presented by R. Agrawal et al.in 1993. Apriori algorithm is expensive to handle a huge number of candidate sets and it requires several scans for the database which is a difficult job. However, in situations with a huge number of frequent patterns, lengthy patterns, or quite little minimum support thresholds, an Apriori*-*like algorithm may undergo from some above problems and it is used for only single dimensional mining.

A. What is Association rule?

Association rules are if /then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository[8].An example of an association rule would be "If a customer buys a toothpaste, he is 80% likely to also purchase toothbrush." An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A subsequent is an item that is found in combination with the antecedent. Association rule is the implication of the form $A \Rightarrow B$, where A and B are item sets which satisfies $A \subseteq I$, $B \subseteq I$ and $A \cap B = \varphi$. In data mining, association rules are useful for analyzing and predicting customer activities. They play significant part in shopping basket data analysis, store layout and product clustering.

Manuscript received July 22, 2014.
Chaitanya B. Pednekar, Computer Engg, Mumbai University/ ACPCE,Kharghar, India. chaitanyap8510@gmail.com
R.C.Suryawanshi, Computer Engg, Mumbai University/ ACPCE,Kharghar, India. rakeshsuryawanshi@gmail.com

B. Association Rule Mining :

in data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is planned to identify strong rules discovered in databases using different measures of interestingness. For example, the rule $buys\{\text{Bread, milk}\} \Rightarrow buys\{\text{Butter}\}$ found in the sales data of a supermarket would indicate that if a customer buys bread and milk together, he or she is likely to also buy Butter. Such information can be used as the basis for decisions about marketing activities such as, promotional pricing. In addition to, the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection. As opposite to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

C. Support and Confidence:

In support-confidence framework, each association rule has support and confidence to confirm the validity of the rule. The support denotes the occurrence rate of an itemset in *DBT* and the confidence denotes the proportion of data items containing *B* in all items containing *A* in *DBT* .

$Sup(i) = Count(i) / Count(DBT)$
$Sup(A \rightarrow B) = Sup(A \cup B)$
$conf(A \cup B) = Sup(A \cup B) / Sup(A)$

II. CLASSIFICATION OF ASSOCIATION RULE

Association rule can be classified based on dimension appearing in the rule. In multidimensional databases we refer each distinct predicate as a dimension.

A. Single dimensional Association Rule:

It contain single distinct predicate with multiple occurrences. That means predicate occur more than once in the rule.eg-$buys(X,\text{"mobile"}) \Rightarrow buys(X,\text{"memory card"})$

B. Multidimensional Association rule:

Association rule that contain two or more dimension or predicates is referred as multidimensional Association rule[4].Each of which occurs only once in the rule so there is no repetitive predicates.
eg.(X,"25..30")∧occupation(X,"student")⟹buys(X,"laptop")

C. Hybrid Dimensional Association Rule:

These are the multidimensional Association rule with repetitive predicates, which contain several occurrences of some predicates.
eg- age(X,"20..30")∧ buys(X,"laptop")⟹ buys(X," b/w printer")

III. APPROACHES FOR MINING ASSOCIATION RULE

There are four approaches for mining association rule. These are as follows:

A. Apriori Algorithm

The traditional Apriori algorithm employs an iterative method to find all the frequent item-sets. First, the frequent 1- item sets L_1 is found according to the user-specified minimum support threshold, and then the L_1 is used to find frequent 2-iemsets L_2, and so on, until there is no new frequent item sets could be found. After finding all the frequent item sets using Apriori, we could generate the corresponding association rules [2]. Apriori employs an iterative approach known as a level-wise search, where k-item sets are used to explore (k+1)-item sets. Apriori principle: If an item set is frequent, then all of its subsets must also be frequent. It works in two steps-Join Step: Ck is generated by joining $L_{k-1}$with itself. Prune Step: Any (k-1)-item set that is not frequent cannot be a subset of a frequent k-item set[8].Apriori Algorithm is the simple Single-dimensional mining algorithm.

B. Partition Algorithm

In this algorithm if we are given a database with a small number of probable large item sets say a few thousands, then support for them can be tested in one scan by using a partitioning technique [10]. Partitioning divides the database into non-overlapping subsets; these are individually considered as separate databases and all large item sets for that partition called local frequent item sets, are generated in one pass. The Apriori algorithm can then be used proficiently on each partition if it fits completely in main memory. Partitions are selected in such a way that each partition can be accommodated in main memory.

C.Sampling Algorithm

The main inspiration for the sampling algorithm[13] is to select small sample one that fits in the main memory of the database of transactions and to determine the repeated item sets from that sample. If those frequent item sets form a superset of frequent item sets for the entire database, then we can determine the real frequent item sets by scanning the rest of the database in order to compute correct support values for the superset item sets. A superset of frequent item sets can generally be found from by using for eg.Apriori algorithm with a lowered minimum support.

D. FP-Growth Algorithm

FP-growth algorithm is a well-organized method of mining all frequent item sets exclusive of candidate's generation. The algorithm mine the frequent item sets by using a divide-and-conquer strategy as follows: FP-growth first compresses the database representing frequent item set into a frequent-pattern tree, or FP-tree, which retains the item set association information as well. The next step is to divide a compressed database into set of restricted databases , each associated with one frequent item. lastly, mine each such database separately. Principally, the construction of FP-tree and the mining of FP-tree are the main steps in FP-growth algorithm.
In actuality, for example, along with items purchased in sales transactional databases, other related information like quantity purchased, branch location, price etc are stored. supplementary related information regarding the customers who purchased the items, such as customer age, occupation, income, and address also stored in the database. Frequent item sets along with other relevant information will be helpful in high-level decision-making. This leads to the challenging mining task of multilevel and multidimensional association rule mining. In recent years, there has been lot of interest in mining databases with multidimensional data values.

IV. CONDITIONAL HYBRID DIMENSIONAL ASSOCIATION RULE MINING

Based on these marking, either it does intra-dimensional join or inter-dimensional join. Thus here we present mining conditional hybrid-dimensional association rules. To solve the problems for finding frequent itemsets we have presented this algorithm. It mines hybrid dimension association rules not only from single-dimensional as well as from multidimensional database. It meets the definite condition to generate conditional hybrid dimensional association rules, from multidimensional transactional database. It scans database only once which makes easy to find large frequent patterns. It does not generate the candidate itemsets as we generate in Apriori algorithm, rather it uses Boolean vector "relational calculus" to generate frequent item sets.
The conditional limit in the rules is, the predicates represent the main attribute can occur many times, but the other predicates refer to subordinate attributes can only occur once [6]. Therefore, to achieve our aim, we design an algorithm for mining hybrid-dimension association rules, which meet the definite condition, from multidimensional transaction database. We focused on finding non-repetitive predicate multi- dimensional rules. We integrate the single-dimensional mining and no repetitive predicate multi-dimensional mining, and present a method for mining hybrid-dimensional association rules using Boolean matrix.

A. The Join Process

There are two steps in generation of the frequent item sets and frequent predicate sets. The two steps are joining and pruning.
1. The join generating candidate 2-itemsets C2:

We find frequent 1-itemsest based on each attribute, at the same time we mark items belong to every main attribute. So it will be clear that the marked items are the items of main attribute and unmarked items are the subordinate items. When we search for ***C2,*** if both of the two joining items are marked items, we call the function for intra-dimensional join between the items as well as inter-dimensional join, but only proceed with inter-dimensional join on the other occasions[2].

2. The join on other occasions: When we generate frequent item sets directly according to the join mode of the Apriori, it would take place intradimensional join as well as inter-dimensional join. But there are some limitations to the generation of intradimensional join and inter-dimensional join. Therefore we make the following modifications to the joining step of the Apriori. We assume that items within transaction and item-set are sorted in lexicographic order. We could take two steps to find L_k[2].

• Distinguish the intra-dimensional join and inter-dimensional join; If all the items within the two (k-1) item-sets belong to the main attribute; we proceed with intra-dimensional join, and proceed with inter-dimensional join on other occasions.

• Implement join $\mathbf{L_{k-1} \bowtie L_{k-1}}$, and choose the corresponding joining condition according to the characteristic of the join (intra-dimensional join or inter-dimensional join)

B. The Conditional Restriction in Hybrid-Dimension Association Rules

First the frequent itemsets are obtained, and then we produce the hybrid-dimension association rules from the frequent item-sets. In the process of generating frequent item-sets, we make both intra-dimensional join and inter-dimensional join, as well as the conditional limitations while proceeding with join, all of the frequent item-sets have such a character: the values within main attribute field occur many times, while the values within subordinate attribute fields occur only once. Thus, the rules generated by the algorithm may include many predicates, or include the same predicate. So the hybrid dimension association rules are formed [2].

V. ALGORITHM

The algorithm consists of following steps:

1. Transforming the multidimensional transaction database into two Boolean matrices [3] one for subordinate attributes (Am*p) and one for main attribute (Bm*q).
2. Generating the set of frequent 1-itemset LA1(from the subordinate attributes matrix) and LB1 (from the main attribute matrix).
3. Pruning the Boolean matrices.
4. Perform AND operations to generate 2-itemsets:
   LA1⋈ LB1 and LA1 ⋈ LA1 for inter-dimension join And LB1⋈ LB1 for intra-dimension join.
5. Repeat the process to generate (k+1)-item-sets from Lk.

A. Transforming the Multidimensional Transaction Database into Boolean Matrix:

At first two separate Boolean matrix is used, one for subordinate attributes and second for main attributes respectively. A Boolean matrix **Am*p** which has m (records or transactions) rows and p (for (n-1) subordinate attributes values) columns and another Boolean matrix **Bm*q** which has m (records or transactions) rows and q (for (nth) main attributes values) columns. For each transaction in subordinate attributes matrix, only one category of each dimension contains the value 1, rest will contain 0.

B. Generating the Frequent 1-itemset L_1

The Boolean matrices **Am*p** and **Bm*q** are scanned and support numbers of all dimension values are computed. The Support number = **Ij** * Sup_num, of **Ij ,** is the number of'1s' in the jth column of the Boolean matrix. If **Ij***sup_num is smaller than min_sup_num, itemset { **Ij** } is not a frequent 1-itemset and the jth column will be deleted from the matrix **.**Otherwise itemset { **Ij** } is the frequent 1-itemset and is added to frequent one itemset list L1.

C. Pruning the Boolean Matrix

Pruning the Boolean matrix means deleting some columns from it.

D. Generating the Set of Frequent k-item Sets Lk

Frequent k-itemsets are discovered by AND relational calculus,which is carried out for the k-vectors combination. If the Boolean matrix Ap*q has q columns where $2<q\leq n$ and *min_sup_num* is $h\leq p\leq m$, $(Cq)^k$, combinations of k-vectors will be produced. The AND relational calculus is for each combination of k-vectors. If the sum of element's values in the 'AND' calculation result is not smaller than the minimum support number *min_sup_num*, the k-itemsets corresponding to this combination of k-vectors are the frequent k-itemsets and are added to the set of frequent k itemsets L_k.

By integrating the single-dimensional mining and non repetitive predicate multi- dimensional mining, a new method is created i.e hybrid- dimensional association rules including Boolean Matrix. Let a multi-dimensional transaction database Order, which includes two subordinate attributes Age & Income and one main attribute Ordered_items as given in table I. In order to simplify the implement process, we pre-processed some attributes before algorithm executes, shown below in table II and table III.

The multidimensional transaction table Order is transformed into two Boolean Matrices: Am*p as subordinate attributes matrix and Bm*q as main attribute matrix. Which are as given below: Let the minimum support is 0.4; m=10 is the number of transactions.

Table 1: Order

ID	Age	Income	Ordered_items
1	31..40	6470	I1, I2, I5
2	31..40	7900	I1, I2
3	31..40	9350	I1, I2, I5
4	0..30	35000	I2, I4
5	41+	17500	I1, I3
6	31..40	8600	I1, I2, I4
7	31..40	3750	I1, I3, I5
8	0..30	25700	I2, I5
9	0..30	28000	I1, I2, I3
10	0..30	30200	I3, I4

Table 2
Mapping Age

Interval	Name
0..30	A
31..40	B
41+	C

Table 3
Mapping Income

Interval	Name
0..15000	P
15001..25000	Q
25001+	R

Table 4
Order Itemsets

ID	Age	Income	Ordered_items
1	B	P	I1, I2, I5
2	B	P	I1, I2,
3	B	P	I1, I2, I5
4	A	R	I2, I4
5	C	Q	I1, I3
6	B	P	I1, I2, I4
7	B	P	I1, I3, I5
8	A	R	I2, I5
9	A	R	I1, I2, I3
10	A	R	I3, I4

Therefore min_sup_num= 4. By computing the sum of the elements value of each column in the Boolean matrix A10*4 and B10*4 set of frequent 1-itemset is:
LA1 = {{A},{B},{P},{R}}, LB1 ={{I1},{I2},{I3},{I5}} smaller than the minimum support number [7]. Now we perform the "AND" operation to join LA1 and LB1 (according to the type of join) to generate L2. The possible 2-itemsets are: Inter-dimensional join (LA1⋈ LB1 and LA1⋈ LA1): It is performed by AND operation among the columns of Matrix Am*p AND Bm*q and Am*p AND Am*p. Intra-dimensional join (LB1⋈ LB1): It is performed by AND operation among the columns of Matrix Bm*p AND Bm*q. The possible 2- itemsets from LA1andLB1are:(a,p),(b,r),(p,1),(p,2),(p,3),(p,5),(r,1),(r,2),(r,3),(r,5),(a,1),(a,2),(a,3),(a,5),(b,1),(b,2),(b,3),(b,5),(I1,I2),(I1,I3),(I1,I5),(I2,I3),(I2,I5),(I3,I5). After performing "AND" operation to get the support numbers of these mentioned item sets the Boolean matrices A10*18 and B10*6 are generated. Now again we compute the sum of the columns of matricesA10*18 and B10*6. And prune the columns of the 2- itemsets those are not frequent. Same process will be repeated till for next higher item sets.
We can generate such a hybrid-dimension association rule: $b \wedge p \wedge I8 \Rightarrow I\ 13$ (Support=40% and Confidence=100%)

VI. EXPERIMENT

To test whether the proposed method is quick, expansible and effective our experiments are made on different operating systems. We test on machine with Intel(R) Core i-3 2328M, 2.20 GHz and 4GB memory. The Operating system is Windows 7. Also we test proposed method on machine Intel Xeon, 2.66 GHz & 4 GB memory. The Operating system is MAC OS X Lion. We use a database that has 300 records and 20 attributes, which have different value. Time value for execution is given in second. We observed that, it shows fraction of time difference on different operating system.

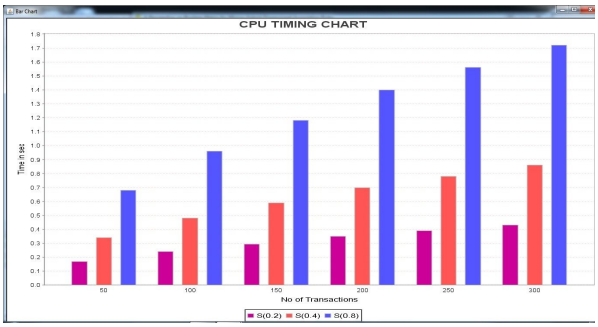

Fig 1. Results obtained on windows 7 operating system

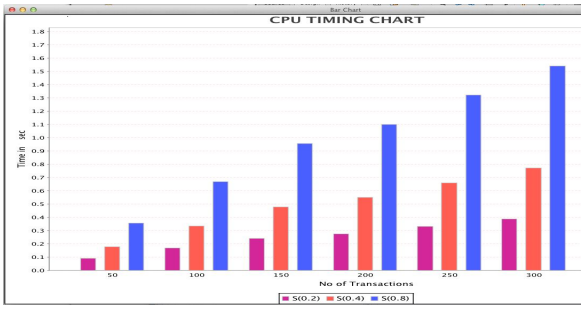

Fig 2. Results obtained on MAC operating system

VIII. CONCLUSION

Proposed system interacts with large database to generate multidimensional association rule. The proposed system is highly scalable and accurate. It scans the database only once and it does not generate the candidate item sets. It uses the boolean vector relational calculus to generate frequent item sets. It stores the data in the form of bits, so it needs less memory space and can be applied to large relational databases. Apriori property is used in algorithm to prune the item sets. It is not necessary to scan the database again; it uses boolean logical operations to generate the association rules.

REFERENCES

[1] R.Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. Intl Conf. Very Large DataBases, pp. 487-499, Sept.1994.

[2] YanXin,Shi-GuangJu,"mining conditional hybrid-dimension association rules on the basis of multi-dimensional transaction database" Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi'an, 2-5 November 2003.
[3] Hunbing Liu and Baishen wang, "An association Rule Mining Algorithm Based On a Boolean Matrix" , Data Science Journal, Volume 6, Supplement 9, S559-563, September 2007.
[4] B. Goethals and M.J. Zaki, "Frequent Item set Mining Implementations" IEEE ICDM Workshop Proc., vol. 80, Nov. 2003.
[5] K. Gouda and M. J. Zaki, "Efficiently mining maximal frequent Itemsets ", In ICDM, pp. 163–170, 2001.
[6] S. Bashir, A. Rauf Baig, "Hybrid Miner: Mining Maximal Frequent Itemsets.A using Hybrid Database Representation Approach", In Proc.of 10th IEEE-INMIC conference, Karachi, Pakistan, 2005.
[7] Jurgen M. Jams Fakultat fur Wirtschafts- irnd, "An Enhanced Apriori Algorithm for Mining Multidimensional Association Rules, 25th Int. Conf. Information Technology interfaces ITI Cavtat, Croatia (1994).
[8] R.Agrawal, H.Mannila, R.Srikant, H.Toivone and A.I.Verkamo, "Fast discovery of association rules" Advances in Knowledge discovery and Data Mining, pages 307–328. AAAI Press, Menlo Park, CA, 1995.
[9] H. Mannila, H. Toivonen, and A. Verkamo. "Efficient algorithm for discovering association rules". AAA1 Workshop on Knowledge Discovery in Databases.
[10] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques". Higher Education Press 2001.
[11] D. Burdick, M. Calimlim, and J. Gehrke, "Mafia: A maximal frequent itemsets: A algorithm for transactional databases", In Proc. of ICDE Conf, pp. 443-452, 2001.
[12] U.M. Fayyed, G.Piatetsky-Sharpiro, P.Smyth,and R.Uthurusamy, editors, "Advances in knowledge Discovery and Data Mining",pages 307-328.AAAI/MIT press, 1996.
[13] H. Toivonen, "Sampling large databases for association rules" In proceeding of the 1996 international conference on Very Large Data Bases(VLDB'96),Bombay,India,pp 134-145,1996.

Mr. C. B. Pednekar received the B.E degree in Computer Science & Engineering from Bharati Vidyapeeth College of Ebgineering, Kolhapur,India,in year 2008. He is doing his M.E in Computer Engineering at Mumbai University, Mumbai,India.He has published various papers in the area of Data Mining.

Prof. R. C. Suryawanshi received the M.E degree in Computer Engineering from Umiversity of Mumbai, India. He is currently working as Associate Professor in Department of Information Technology,ACPCE,Kharghar,Mumbai,India.He has published various papers in the area of System Security & Data Mining.