Survey on Multidimensional and Conditional Hybrid Dimensional Association Rule Mining

Nilam K. Nakod, M. B.Vaidya

Abstract— Association rule mining is important research topic today. In this paper I have presented the overall survey of multidimensional association rule as well as the survey of Hybrid dimensional association rule mining .This paper illustrate the different approaches for mining multidimensional as well as hybrid dimensional association rule. This paper also elaborates the conditional hybrid dimensional association rule and concludes which is the best Approach for mining the multidimensional and conditional hybrid dimensional Association rule.

Index Terms—Boolean Relational Calculus, multi-dimensional Association Rule, multi-dimensional Transactional Database, RSHAR..

I. INTRODUCTION

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[5]. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk."
An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.
In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, and catalog design and store layout.

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 intended to identify strong rules discovered in databases using different measures of interestingness.

For example, the rule

$$buys\{Onions, Potatoes\} \Longrightarrow buys\{Burger\}$$

found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. 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 and bioinformatics. As opposed 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-

$$buys(X, \text{"Digital camera"}) \Longrightarrow buys(X, \text{"HP Printer"})$$

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-

$$age(X, \text{"19..24"}) \wedge occupation(X, \text{"}student\text{"}) \Longrightarrow buys(X, \text{"}laptop\text{"})$$

C. Hybrid Dimensional Association Rule:

These are the multidimensional Association rule with repetitive predicates, which contain multiple occurrences of some predicates.eg-

$$age(X, \text{"19..24"}) \wedge buys(X, \text{"}laptop\text{"}) \Longrightarrow buys(X, \text{" }b/w\ printer\text{"})$$

Manuscript received February 2013.
Nilam K.Nakod, Computer, Pune / AVCOE, City Sangamner, India,
M.B.Vaidya, Computer, University/ College,AVCOE,Sangamner,India.

III. APPROACHES FOR MINING ASSOCIATION RULE

A. Apriori Algorithm

The classical Apriori algorithm employs an iterative method to find all the frequent item-sets. First, the frequent 1- item sets L1 is found according to the user-specified minimum support threshold, and then the L1 is used to find frequent 2-iemsets L2, 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[5]. 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 Lk-1with itself. Prune Step: Any (k-1)-item set that is not frequent cannot be a subset of a frequent k-item set. Apriori Algorithm is the simple Single-dimensional mining algorithm.

B. Sampling Algorithm

The main idea for the sampling algorithm[3] is to select small sample one that fits in the main memory of the database of transactions and to determine the frequent 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 remainder of the database in order to compute exact support values for the superset item sets. A superset of frequent item sets can usually be found from by using for eg. Apriori algorithm with a lowered minimum support.

C. Partition Algorithm

The Partition algorithm differs from the Apriori algorithm in terms of the number of database scans[10].In this algorithm if we are given a database with a small number of potential large item sets say a few thousands, then support for them can be tested in one scan by using a partitioning technique. 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 efficiently on each partition if it fits entirely in main memory. Partitions are chosen in such a way that each partition can be accommodated in main memory.

D. FP-Growth Algorithm

FP-growth algorithm is an efficient method of mining all frequent item sets without candidate 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 conditional databases (a special kind of projected database), each associated with one frequent item. Finally, mine each such database separately. Particularly, the construction of FP-tree and the mining of FP-tree are the main steps in FP-growth algorithm.

In reality, for example, along with items purchased in sales transactional databases, other related information like quantity purchased, price, branch location etc are stored. Additional related information regarding the customers who purchased the items, such as customer age, occupation, credit rating, 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. APPROACHES FOR MINING MULTIDIMENSIONAL ASSOCIATION RULE

A. First Approach:

Quantitative attributes are discretized using predefined concept hierarchies. This discretization occurs prior to mining. For instance, a concept hierarchy for income may be used to replace the original numeric values of this attribute by ranges, such as “0-20K", “21-30K", “31-40K", and so on. Here, discretization is static and predetermined. The discretized numeric attributes, with their range values, can then be treated as categorical attributes (where each range is considered a category). We refer to this as mining multidimensional association rules using static discretization of quantitative attributes.

B. Second Approach :

Quantitative attributes are discretized into \bins" based on the distribution of the data. These bins may be further combined during the mining process. The discretization process is dynamic and established so as to satisfy some mining criteria, such as maximizing the confidence of the rules mined. Because this strategy treats the numeric attribute values as quantities rather than as predefined ranges or categories, association rules mined from this approach are also referred to as quantitative association rules.

C. Third Approach :

Quantitative attributes are discretized so as to capture the semantic meaning of such interval data. This dynamic discretization procedure considers the distance between data points. Hence, such quantitative association rules are also referred to as distance-based association rules.

D. Boolean Matrix based Approach for mining multidimensional Association Rule:

In this, algorithm based on Boolean matrix[9] are used to generate the multidimensional rule which has no repetitive predicates. A Boolean Matrix based approach has been used to find the frequent itemsets, the items forming a rule come from different dimensions. It is an algorithm for mining multidimensional association rules from relational databases. The algorithm adopts Boolean relational calculus to discover frequent predicate sets. When using this algorithm first time, it scans the database once and will generate the association rules. 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. It stores all data in the form of bits, so it needs less memory space and can be applied to large relational databases.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. CONDITIONAL HYBRID-DIMENSION ASSOCIATION RULES MINING

The generation of frequent itemsets is the core of all the association rules mining algorithms. Previous studies on mining multi-dimensional association rules we focused on finding non-repetitive predicate multi- dimensional rules.
We integrate the single-dimensional mining and non-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 itemsets 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 f intra-dimensional join between the items as well as inter dimensional join, but only proceed with inter-dimensional join on the other occasions.

2) The join on other occasions

When we generate frequent itemsets directly according to the join mode of the Apriori, it would occur intra-dimensional join as well as inter-dimensional join. But there are some restrictions to the generation of intra-dimensional 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 itemset are sorted in lexicographic order. We could take two steps to find L_k

a) Distinguish the intra-dimensional join and inter-dimensional join

If all the items within the two (k-1)- itemsets belong to the main attribute; we proceed with intra-dimensional join, and proceed with inter-dimensional join on other occasions. Implement join $\boldsymbol{L_{k-1}} \bowtie \boldsymbol{L_{k-1}}$, and choose the corresponding joining condition according to the characteristic of the join (intra-dimensional join or inter-dimensional join) [6].

b) The conditional restriction in hybrid-dimension association rules

First the frequent itemsets are obtained, and then we generate the hybrid-dimension association rules from the frequent itemsets. In the process of generating frequent itemsets, we make both intra-dimensional join and inter-dimensional join, as well as the conditional restrictions while proceeding with join, all of the frequent itemsets 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.

VI. APPROACHES FOR MINING HYBRID DIMENSIONAL ASSOCIATION RULES

A. Rough set Model:

In this model, the mining of hybrid association rules using rough set approach[7].This algorithm can be called as RSHAR. The RSHAR algorithm is constituted of two steps mainly. At first, to join the participant tables into a general table to generate the rules which is expressing the relationship between two or more domains that belong to several different tables in a database. Then we apply the mapping code on selected dimension, which can be added directly into the information system as one certain attribute. To find the association rules, frequent itemsets are generated in second step where candidate itemsets are generated through equivalence classes and also transforming the mapping code in to real dimensions. The searching method for candidate itemset is similar to apriori algorithm. The analysis of the performance of algorithm has been carried out.

B. Novel Algorithm:

Association rule mining is a fundamental and important functionality of data mining. Most of the existing real time transactional databases are multidimensional in nature. a novel algorithm is proposed for mining hybrid-dimensional association rules which are very useful in business decision making. This algorithm uses multi index structures to store necessary details like item combination, support measure and transaction IDs, which stores all frequent 1-itemsets after scanning the entire database first time. Frequent k-itemsets are generated with previous level data, without scanning the database further. Compared to traditional algorithms, this algorithm efficiently finds association rules in multidimensional datasets, by scanning the database only once, thus enhancing the process of data mining.

C. Data cube:

The mining of single-dimensional association rule and non-repetitive predicate multi-dimensional association rule were integrated. It uses the data cube structure [8] for mining hybrid dimensional association rule. Mining single variable hybrid-dimension association rules. It does not mine the multi-variable hybrid-dimension association rules.

D. Boolean Matrix approach for mining conditional hybrid dimensional Association Rule:

In this, algorithm based on Boolean matrix [9] are used to generate the multidimensional rule with repetitive predicates. That is for mining the conditional Hybrid dimensional association rule. All other approaches are not suitable for conditional hybrid dimensional Association Rule mining. A Boolean Matrix based approach has been used to find the frequent itemsets, the items forming a rule come from different dimensions. It is an algorithm for mining Conditional Hybrid dimensional Association Rule from multidimensional Transaction Databases. The algorithm adopts Boolean relational calculus to discover frequent predicate sets. When using this algorithm first time, it scans the database once and will generate the association rules. 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.

It stores all data in the form of bits, so it needs less memory space and can be applied to large relational databases.

VII. CONDITIONAL HYBRID-DIMENSION ASSOCIATION RULES MINING

The generation of frequent itemsets is the core of all the association rules mining algorithms. Previous studies on mining multi-dimensional association rules we focused on finding non-repetitive predicate multi- dimensional rules.
We integrate the single-dimensional mining and non 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 itemsets 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 f intra-dimensional join between the items as well as inter dimensional join, but only proceed with inter-dimensional join on the other occasions.

2) The join on other occasions

When we generate frequent itemsets directly according to the join mode of the Apriori, it would occur intra-dimensional join as well as inter-dimensional join. But there are some restrictions to the generation of intra-dimensional 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 itemset are sorted in lexicographic order. We could take two steps to find L_k

a) Distinguish the intra-dimensional join and inter-dimensional join

If all the items within the two (k-1)- itemsets belong to the main attribute; we proceed with intra-dimensional join, and proceed with inter-dimensional join on other occasions. Implement join $\boldsymbol{L_{k-1} \bowtie L_{k-1}}$, and choose the corresponding joining condition according to the characteristic of the join (intra-dimensional join or inter-dimensional join) [6].

b) The conditional restriction in hybrid-dimension association rules

First the frequent itemsets are obtained, and then we generate the hybrid-dimension association rules from the frequent itemsets. In the process of generating frequent itemsets, we make both intra-dimensional join and inter-dimensional join, as well as the conditional restrictions while proceeding with join, all of the frequent itemsets 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

VIII. CONCLUSION

In This paper I have Presented the overall survey of mining multidimensional as well as the survey of mining the conditional Hybrid dimensional Association Rule mining. From this comparative study, the Boolean matrix based Approach is best suited for mining multidimensional Association Rule And also for mining conditional Hybrid dimensional Association Rule. It can mine the multidimensional Association rule from Relational Database and it can also mine the conditional Hybrid Dimensional Association Rule from multidimensional transactional Database. A Boolean Matrix based approach has been used to find the frequent itemsets, the items forming a rule come from different dimensions. It is an algorithm for mining Conditional Hybrid dimensional Association Rule from multidimensional Transaction Databases. The algorithm adopts Boolean relational calculus to discover frequent predicate sets. When using this algorithm first time, it scans the database once and will generate the association rules. 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. It stores all data in the form of bits, so it needs less memory space and can be applied to large relational databases.

ACKNOWLEDGMENT

The author would like to thank “Amrutvahini College of Engineering, Sangamner.” and Prof. M. B. Vaidya for their help and suggestions. We would also like to thank International Journal of Emerging Science and Engineering (IJESE), for publishing paper.

REFERENCES

1. Shenoy, P.”Turbo-charging Vertical Mining of Large Databases”. In ACM SIGMOD International Conference on Management of Data. Dallas. 2000.
2. Anjana P, Kamalraj P,”Rough set model for discovering Multidimensional association rules”, in the proceedings of IJCSNS, VOL 9, no.6, p p159-164, 2009
3. 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.
4. Srikant R, Agrawal R “Mining quantitative association rules in large relational table”, in the proceedings of ACM SIGMOD international conference on management of data, p.1-12 1996.
5. Agrawal R,. Mannila H,. Srikant R,. Toivonen H, and Inkeri Verkamo A. “Fast discovery of association rules”. Advances in Knowledge Discovery and Data Mining, pages 307–328. AAAI Press, Menlo Park, CA, 1995
6. Yan Xin, Shi-Guang Ju , “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.
7. Anjna Pandey,K.R.Pardasani,”Rough set Model for Discovering Hybrid Dimensional Association Rules”,International Journal of Computer Science and Network security,Vol 9,No.6,pp151-164, 2009
8. ZHi-jie LI,”Using Data Cube for Mining Hybrid Dimensional Association Rule,”GCC 2003 LCNS3033 pp-899-902 , 2004 @ Springer – Verlag Berlin Heidelberg 20004
9. Hunbing Liu and Baishen Wang,”An Association Rule Mining Algorithm Based On Boolean Matrix”,Data Science Journal,Volume 6,Supplement 9,S559-563,September 2007
10. Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques”. Higher Education Press 2001.

.