ISSN No. 0976-5697

Volume 2, No. 4, July-August 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Mining Multidimensional Association Rules

Rakesh Sharma*
Department of Information Technology
Haryana College of Technology & Management,
Kaithal, India
rakeshsharma3112@gmail.com

Pinki Sharma
Department of Computer Science & Engineering
Haryana College of Technology & Management,
Kaithal, India
pinkisharma@gmail.com

Abstract: Association rule mining plays an important role in knowledge and information discovery. However, existing association rule mining is only focused on single level datasets. In this paper, we firstly present a introduction for multidimensional association rules, then we present introduction about static discretization approach for mining multidimensional association rule In previous studies the association rules are generated as single dimension however mining association rules at multiple dimension may lead to the discovery of more specific and concrete knowledge from large transaction databases by extension of some existing rules mining techniques. In multidimensional association rules we use same minimum support for different conceptual levels. In this paper, we also discover multidimensional (cross–level) association rules using MLT2 algorithm detailed in Han and Fu's paper. This algorithm discovers association rules for successive levels making use of rules already discovered for cross levels of concept hierarchy.

Keywords: Data mining, support, Association rules, Multidimensional Association rule.

I. INTRODUCTION

Data mining is the technique of extracting in formation from large observational data banks i.e. mostly unorganized. It is the process that refers to performing automated extraction and generating predictive information from large data sets. The typical example of a data mining problem is "market basket analysis". Stores maintain information on what transactions are carried by their customers. By finding out what products are frequently purchased jointly (i.e. are associated with each other), being able to optimize the marketing of the products (e.g. the layout of the store) by better targeting certain crowds of customers. Association rule mining is a technique for discovering unsuspected data dependencies and is one of the best known data mining techniques. The use of association rules enables to ascertain union and relationships between large unspecified data items based on certain attributes and characteristics. Association rule mining thus solves the problem of how to search efficiently for those associations and relationships [1, 2, 6]. Association rules can be categorizes in a variety of ways, based on the following criteria:

Boolean and quantitative association rule: The rules that are only concern about the presence or absence of items are known as Boolean association rule if a rule defines associations between quantitative items or attributes, then it is a quantitative association rule. In these rules, quantitative values for items or attributes are divided into ranges.

Single dimensional and Multidimensional association rules: If the items or attributes in an association rule not considering any dimension and considering all items at one dimension, then it is a single-dimensional association rule. If a rule indicates two or more dimensions, then it is a multidimensional association rule.

Single level and multilevel association rule: when finding association rule item are not considering any abstract level then it is called single level association rule and when we consider items at various levels of abstraction then we called such type of rule multilevel association rules.

Based on the nature of the association involved in the rule: Association mining can be extended to correlation analysis, where the absence or presence of correlated items can be identified[1].

We are only concern with multidimensional association rules because in some cases it might be useful to discover rules among items from different levels of the concept hierarchy For many applications; it is difficult to find strong associations among data items at low or primitive levels of abstraction due to the sparsity of data in multidimensional space. Strong associations discovered at very high concept levels may represent common sense knowledge. However, what may represent common sense to one user may seem novel to another. Therefore, data mining systems should provide capabilities to mine association rules at multiple levels of abstraction and traverse easily among different abstraction spaces.

II. MULTIDEMENSIONAL ASSOCIATION RULES

Most of the algorithms and techniques only concern about association rules within single attribute and Boolean data, all those rules are about the same attribute and the value can only be *yes/1*or *no/0*. By mining multiple dimensional association rules we can generate the rules such as:

age(X, "19 – 24") ^ buys(X, "laptop")) buys(X, "b/w printer")

Multiple dimensional association rule mining is to discovery the correlation between different predicts/attributes. A dimension can be a attribute or a predicate, such as: age, occupation and buys in the given example. Multiple dimensional association rule mining concerns data of type such as Boolean data, categorical data and numerical data [1,2,3] The process of mining multiple dimension association rules is similar to the process of mining any multiple level association rule mining. Firstly

generating 1-dimensions frequent itemsets after that generating all frequent itemsets based on any single level itemset generation algorithm. The mining process is simple; however three basic approaches are used for generating multiple dimensions. One of the approaches is u*sing static discretization method*, in this method, quantitative attributes are separated into different ranges according the defined hierarchies and attributes are replaced by separated ranges earlier to the mining process. Categorical attributes also can be generalized to higher concept level if necessary. After this process the task relevant data can be stored in the table.

The relevant data can also be stored in the data cube, which is more suitable for multiple dimensional association rules since data cube itself is multidimensional by definition. The Apriori[5] can be easily adapted to get the frequent k-predicts by searching through all the relevant attributes instead of one attribute only. The second approach uses d*ynamical discretization method*. In this method quantitative attributes are dynamically discretized during the mining process so as to satisfy some minimum support. Mining quantitative association rules was introduced by Agrawal [3]. Another approach uses *distance based discretization*. In this method distance is used to mining quantitative attributes instead of the concept hierarchy method for each attribute.

In this approach values that are close together are clustered into the same interval. There are two steps, firstly clustering technique is used to generate clusters or intervals, and then distance based association rules are generated by probing for groups or clusters that occur frequently together. The associated parameters were used as metrics during the rule generation process. Those parameters include density and frequency. In this case rules that satisfy certain minimum density and frequency thresholds are taken as interesting rules [1, 3].

III. MINING MULTI DEMENSIONAL ASSOCIATION RULES

As discussed in section 2 there are three approaches mining multidimensional association rules. Our main focus on the *static* discretization approach *for* mining multidimensional association rule. Discritization is a process that transforms quantitative data into qualitative data for example, quantitative data represented by the attributes age in numerical values are represented in descriptive terms such as young and old. It divides the value range of the quantitative attribute into finite number of intervals .The mapping function associates all the quantitative values in a single interval to a single qualitative value. A cut point is a value of the quantitative attribute where a mapping function locates an interval boundary. Diverse taxonomies exist in literature to categorize discretization methods these are complimentary, each relating to a different dimension along which discretization methods may differ.typically discretization without refernce to any other discretization method. One such method is hierarchical discretization utilizes an incremental process to select cut points. This creates an implicit hierarchy over the value ranges. Hierarchcal discretization can be further characterized as either split or merge. Split discretization starts with a single interval that encompasses the entire value range, then repeatedly splits it into sub interval value in a separate interval, then reputedly merges adjacent interval until a stopping criterion is met.As shown in figure.

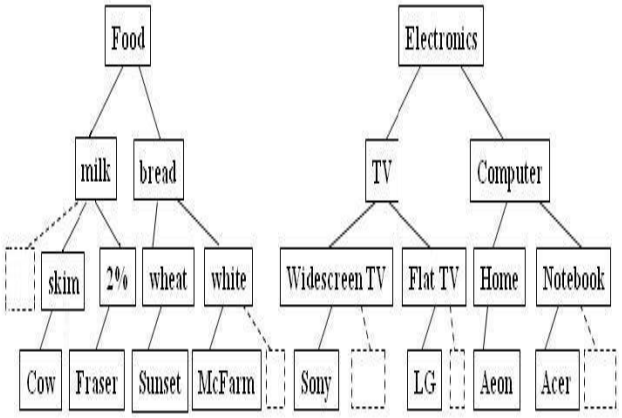

Figure 1 concept hierarchy

An example of multiple dimensional rule from the concept hierarchy illustrated in Figure 1 may be: *Milk* → *sunset [Bread – wheat]*.in this rule milk is upper level and sunset wheat bread is at lower levels of hierarchy generating a cross level association.

A. Multidemension Association Mining Algorithm

MLT2 is basic multiple association rule mining algorithm that generates itemset at same concept level and each level have different support[4]. This algorithm can be used to find multiple dimensional association rules by slightly modifying it and considering same support for all the levels

Following is the algorithm for ML-T2 in pseudo code

for (l:=1;$L[l,1]$ <> 0 and l < max_level; l++) do begin
if l = 1 then begin
$L[l,1]$:= *get_large_1_itemsets* ($T[1],l$);
$T[2]$:= *get_filtered_transaction_table* ($T[1]$, $L[1,1]$);
end
else $L[l,1]$:= get_large_1_itemsets ($T[2]$, l);
for (k := 2; $L[l, k-1]$ <> 0; k++) do begin
C_k := *get_candidate_set* ($L[l, k-1]$);
foreach transaction t in $T[2]$ do begin
C_t := *get_subsets* (C_k , t);
do c.support++;
end
$L[l,$ k] := {c in C_k | c.support >= minsup[l]}
end
$LL[l]$:= U_k $L[l, k]$;
end

This algorithm takes two inputs. First one is hierarchical encoded data where encoding refers to the process of specifying levels to each item in the concept hierarchy. The transaction table represents the data where each instance in the dataset represents one transaction of the form - record id, itemset. Itemset is the list of items for that record id. Second, the minimum support is represented in the algorithm as minsup[l] for each concept level l i.e. uniform for all levels. Our goal is to generate frequent itemsets from this algorithm for mining strong cross-level association rules[7].

IV. ASSOCIATION RULE GENERATION USING MLT2

The results of the algorithms have been checked upon the two different datasets of different size (different number of items and different no of transactions). This data is available at UCI repository [8]. The support factor is changed while taking the results. The following are the basic parameters for analyzing: (1) the number of frequent itemsets generated (2) The execution time (3) The minimum support threshold. The datasets are given as follows.

Table 1 Dataset Table

S.No	Dataset Name	Size Dataset	No. of Transactions
1.	DB1(Soybean)	179KB	980
2.	DB2(Credit-g)	114KB	1000

Summary of Results using Min_Support and No. of Frequent Itemsets Generated Factor are given below:-
On the basis of Min_support and frequent itemset the following graphs (figure 2 and 3), based on the tables can be drawn for analyzing the results. The graphs are as follows.

Parameters: DB1

Table2. Support vs. item set generation for DB1

Min_support	Itemset Generated	
	Level 1	Level2
0.1	87	80
0.2	63	56
0.3	47	24
0.4	23	16
0.5	15	14
0.6	15	8
0.7	15	8
0.8	15	0
0.9	7	0
1	7	0

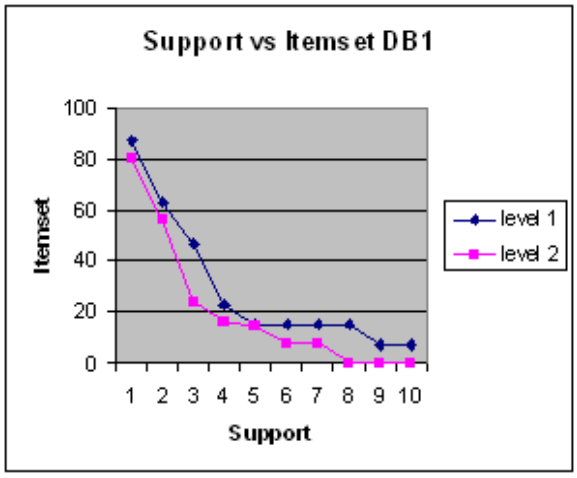

Figure 2: Graph 1

Parameters: DB 2

Table3. Support vs. item set generation for DB2

Min_Support	Itemset Generated	
	Level 1	Level2
0.1	199	112
0.2	136	16
0.3	86	16
0.4	59	4
0.5	39	0
0.6	31	0
0.7	31	0
0.8	23	0
0.9	7	0
1	1	0

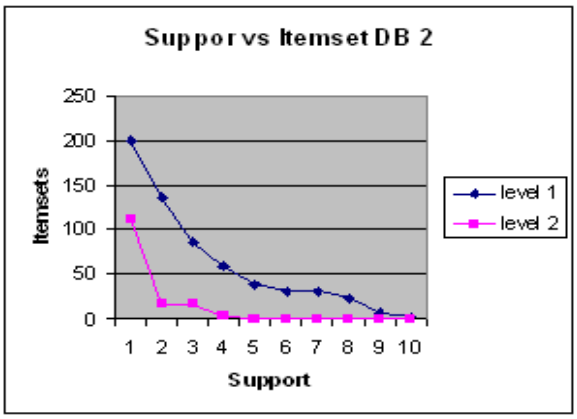

Figure 3: Graph 2

As Min_support decreases at lower levels, the user got more specific information. The generations of frequent item sets as multidimensional are greater than single dimension. More specific information for the users due to reduced support at lower levels

Parameters: DB, DB 2

Table4. Execution time for DB1, DB2

Min_Support	Execution time	
	DB 1	DB 2
0.1	90	95
0.2	84	90
0.3	71	85
0.4	65	75
0.5	50	70
0.6	42	44
0.7	37	33
0.8	29	25
0.9	24	24
1	21	16

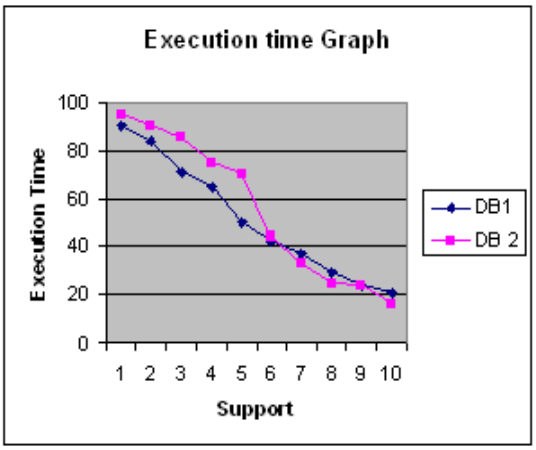

Figure 4: Graph 3

It is clear from above results that as Min_support decreases the execution time of the algorithm is increase (figure 4). The execution time of the algorithm is variable for different datasets with a variation in Min_Support. The time for different frequent item set mining algorithms depends a lot on the structure of the data set

V. CONCLUSION

This study demonstrates that mining multidimensional knowledge is both practical and desirable. This work has successfully discovered multidimensional association rules using MLT2 algorithm [4]. The association rules discovered provides more specific information for the users at cross level of abstraction. Algorithm has efficiently discovered multidimensional association rules from two datasets (credit-g, soybean) from UCI repository [8]. We have noticed that the execution time of the algorithm depends on the size and complexity of concept hierarchy discovered and hence it is variable for different datasets. Algorithm discovers association rules for successive levels making use of rules already discovered for upper levels of concept hierarchy. Number of association rules discovered depends on value of parameters at each level like support, confidence, and lift.

This work is contribution towards representing knowledge at multi dimensional in the form of association rules that enhances the ease and comprehensibility of the users.

VI. ACKNOWLEDGMENT

In the field of association rule mining, most of the proposed methods for generating frequent patterns use the Apriori algorithm.

Two main disadvantages to the Apriori approach are: First, the method may need to generate a large number of candidate sets. Second, repeated scans of the dataset to match and tally the patterns of candidates can be potentially time consuming where the dataset is large and/or the item set is large. So any algorithms that generate item set without candidate generation can be used to make item generation more effective [5].

VII. REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining: Concept and Tech-niques," 3nd Edition, 2006,ch 5.

[2] Fu, Yongjlan, "Data Mining: Tasks, techniques and applications," IEEE. Potentials, 1997, pp.18-20.

[3] R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", SIGMOD 1996

[4] Jiawei Han and Yongjian Fu, "Discovery of Multiple-Level Association Rules from Large Databases," in Proc. 21st VLDB Conference, 1995.

[5] R. Agrawal and R, Shrikanth, "Fast Algorithm for Mining Association Rules". Proceedings Of VLDB conference, Santigo, Chile, 1994, pp 487 – 449.

[6] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases". In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, May 26-28 1993, pp 207-216.

[7] https://www.mscs.mu.edu/~cstruble/class/mscs228/fall2003/project/group4/files/FinalReport.doc

[8] http://repository.seasr.org/Datasets/UCI/arff