Rough Set Model for Discovering Multidimensional Association Rules

Anjana Pandey and KamalRaj Pardasani

Deptt of Information Technology, Deptt of Mathematics

UIT,RGPV Bhopal,India , Maulana Azad National Institute of Technology,Bhopal,India

Summary

In this paper, the mining of multidimensional association rules with rough set approach is investigated as the algorithm RSMAR.The RSMAR 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.

Key words:

Rough Set, multidimensional, inter-dimension association rule, data mining.

1. Introduction

Association rule mining finds interesting association or correlation relationship among a large data set of items [1, 2]. The discovery of interesting association rules can help in decision making process.Association rule mining that implies a single predicate is referred as a single dimensional or *intradimension association rule* since it contains a single distinct predicate with multiple occurrences (the predicate occurs more than once within the rule). The terminology of *single dimensional or intradimension association rule* is used in multidimensional database by assuming each distinct predicate in the rule as a dimension. For instance, in *market basket analysis*,. In *market basket analysis*, it might be discovered a Boolean association rule "laptop b/w printer" which can also be written as a single dimensional association rule as follows [3]:

Rule-1

buys(*X*, "laptop") *buys*(*X*, "b/w printer"),

where *buys* is a given predicate and *X* is a variable representing customers who purchased items (e.g. *laptop* and *b/w printer*). In general, *laptop* and *b/w printer* are two different data that are taken from a certain database

attribute, called *items*. In general, *Apriori* [1] is used an influential algorithm for mining frequent itemsets for generating Boolean (single dimensional) association rules.

Additional relational information regarding the customers who purchased the items, such as customer age, occupation, credit rating, income and address, may also have a correlation to the purchased items. Considering each database attribute as a predicate, it can therefore be interesting to mine association rules containing *multiple* predicate, such as:

Rule-2:

Age ("20..29") ? sex("Male") ? income("5K..7K") ? buys("Laptop")

Where there are four predicates, namely age, sex,income and buys. Association rules that involve two or more dimensions or predicates can be referred to as *multidimensional association rules*. Multidimensional rules with no repeated predicates are called *interdimension association* rules (e.g Rule-2)[4] .On the other hand, multidimensional association rules with repeated predicates, which contain multiple occurrences of some predicates, are *called hybrid-dimension association rules* .The rules may be also considered as combination (hybridization) between intradimension association rules and interdimension association rules. An example of such a rule is the following, where the predicate *buys* is repeated

Rule-3

Age ("20..29") ? sex("Male") ? income("5K") ? buys("Laptop") ? buys("Laser Printer")

Here, we may firstly interested in mining multidimensional association rules with no repeated predicates or interdimension association rules. Hybrid dimension association rules as an extended concept of

Manuscript received June 5, 2009
Manuscript revised June 20, 2009

multidimensional association rules will be discussed later in our next paper. The interdimension association rules may be generated from a relational database or data warehouse with multiple attributes by which each attribute is associated with a predicate. Conceptually, a multidimensional association rule, consists of *A* and *B* as two datasets, called Condition and decision, respectively.

The structure of the paper is the following. Section 2 describes data preparation for the further process of generation rules. Here we will discuss a process of joining table from database.Afterthat relational schema has to be transformed into bitmap table. Section 3, presents the rough set model which is used in RSMAR.Section 4, introduces RSMAR algorithm for mining of interdimension association rules with rough set. Section 5 presents some performance result showing the effectiveness of our method. Finally, section 6 concludes the paper.

2. Background

In this section we provide a short introduction of process of joining tables from relational database and concept of bitmap table which are used in our algorithm.

2.1 Method for joining of Tables

In general, the process of mining data for discovering association rules has to be started from a single table (relation) as a source of data representing relation among item data. Formally, a relational data table [5] *R* consists of a set of tuples, where t_i represents the ^{D_j} *i*-th tuple and if there are *n* domain attributes *D*, then $t_i = (d_{i1}, d_{i2}, \cdots, d_{in})$. Here, d_{ij} is an atomic value of tuple t_i with the restriction to the domain where $d_{ij} \in D$. Formally, a relational data table *R* is defined as a subset of the set of cross product , $D_1 \times D_2 \times \cdots \times D_n$.

where $D = \{D_1, D_2 ....., D_n\}$

Tuple *t* (with respect to *R*) is an element of *R*. In general, *R* can be shown in Table 1

Table1: A Relational Database

Tuples	D_1	D_2	$\cdots$	D_n
t_1	d_{11}	d_{12}	$\cdots$	d_{1n}
t_2	d_{21}	d_{22}	$\cdots$	d_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_r	d_{r1}	d_{r2}	$\cdots$	d_{rn}

In many case the database may consist of several relational data tables in which they have relation one to each others. Their relation may be represented by Entities Relationship Diagram (ERD). Hence, suppose we need to process some domains (columns) data that are parts of different relational data tables, all of the involved tables have to be combined (joined) together providing a *general data table.* In the process of joining the tables, it is not necessary that all domains (fields) of the all combined tables have to be included in the targeting table. Instead, the targeting table only consists of interesting domains data that are needed in the process of mining rules. The process of joining tables can be performed based on two kinds of data relation as follows.

- On the basis of Metadata

Information of relational tables can be stored in a metadata. Simply, a metadata can be represented by a table. Metadata can be constructed using the information of relational data by an Entity relationship Diagram (ERD). A detailed description of metadata and ERD can be found in inten[6]).

- On the basis of function defined by the user.

It is possible for user to define a mathematical function (or table) relation for connecting two or more domains from two different tables in order to perform a relationship between their entities. Generally, the data relationship function performs a mapping process from one or more domains from an entity to one or more domains from its partner entity. Four possibilities of function f performing a mapping process are given by [6]

- One to one relationship

$$f : C_i \rightarrow D_k$$

- One to many relationship

$$f : C_i \rightarrow D_{p1} \times D_{p2} \times ..... \times D_{pk}$$

- Many to one relationship

$$f : C_{m1} \times C_{m2} \times ..... \times C_{mk} \rightarrow D_k$$

- Many to many relationship

$$f : C_{m1} \times C_{m2} \times ..... \times C_{mk} \rightarrow D_{p1} \times D_{p2} \times ..... \times D_{pk}$$

2.2 Data Structure 'Bitmap'

In relation table some attributes has quantitative values which can be discretized as some categorical values on behalf of certain range. Then the form of information system is changed to that each attribute in the new database is an exact value of one item in original system, and each attribute value is either 1 or 0,expressing if it is present there is a '1', otherwise a '0' in the bitmap[7].

Example 1. For an attribute with no- binary domain. each attribute value corresponds to one item. for example, for attribute 'age' with domain(age)={young,middle,old} (i={1,2,3} the following items result:A_1 ="age_young",A_2="age_middle",A_3="age_old" (see fig.1)

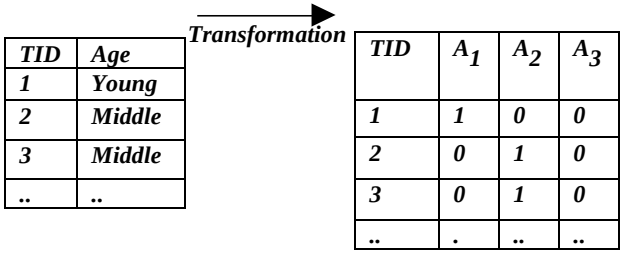

Fig 1.Transformation of relational data into an efficient bitmap representation for attributes with no-binary domains.

3. Rough set

In 1982 Z.Pawalak [8] introduced a new tool to deal with vagueness, called the "rough set". It is a method for uncovering dependencies in data, which are recorded by relations. The rough set philosophy is based on the idea of classification. A detailed introduction to rough set theory can be found in Munakata [9] .

3.1 Model

The rough set method operates on data matrics, so called "Information System.It contains data about the universe U of interest, condition attributes and decision attributes. The goal is to derive rules that give information how the decision attributes depend on the condition attributes. By an information system S, S= {U, *At*, *V*, *f*}, where U is a finite set of objects, U= { $x_1, x_2, ... x_n$ }, *At* is a finite set of attributes, the attribute in *At* is further classified into two disjoint subsets, condition attributes C and decision attribute D, $At = C \cup D$ where C= $c_1 \wedge c_2 \wedge c_3 ..... \wedge c_n$ and D= $d_1 \wedge d_2 \wedge d_3 ..... \wedge d_n$

V= $\bigcup_{p \in A} V_p$,and V_p is a domain of attribute p.

The function f performs a mapping code of condition attributes such that $c_1, c_2 \ldots c_n$ into one simple attribute C which can be added directly into the information system as one certain attribute, it will only posses one column in the information system, analogous
an item.

$f : U \times At \rightarrow V$ is a total function such that $f(x_i, q) \in V_q$.

A prerequisite for rule generation is a partitioning of U in a finite number of blocks, so called equivalence classes[10], of same attribute values by applying an equivalence relation. for example the equivalence Relation $R_1 = \{(u, v, w) \mid u(age, sex, income)\}$ leads to a partition of U into three equivalence classes U_1 ={Adams,Brown}, U_2 ={Carter} and U_3 ={Ford, Gill}(see table 2).Given these classes, rules like e.g. "If age young,sex male and income high then he buys laptop" can be derived . Generally, no unique rule derivation is possible. For example, *Ford* and *Gill* have identical values of the condition attributes, but differ in their values of the decision attribute. In order to analyze such data, the concept of approximation spaces[11] is used to determine the confidence of the derived rules. The quality of the extracted association rules depends also strongly on the possibility of attribute reduction. Reducing the number of attributes in a dataset by removing the redundant ones is one of the main objectives of rough set theory and at the same time one of the main problems. In our approach we try to reduce the computing time by applying the concept of reduct extraction directly to the produced rules, not to attributes. First, in order to generate strong rules, all rules which support and confidence values don't reach the given minimum threshold, are deleted. Second, the reducts are extracted. Suppose, there are two rules with same decision item and same confidence value, and their only difference

is the set of condition items. The condition itemset of rule no 1 is a subset of the condition itemset of rule no 2. In this special case rule no 2 is redundant and can be deleted without having loss of information.

Table 2.Information Table

Universe	*Condition attributes*			*Decision attribute*
Person	*Age*	*Sex*	*Income*	*Buys(Laptop)*
Adams	*Young*	*M*	*High*	*Yes*
Brown	*Young*	*M*	*High*	*Yes*
Carter	*Young*	*M*	*Low*	*No*
Ford	*Middle*	*M*	*High*	*Yes*
Gill	*Middle*	*M*	*High*	*No*

4. Proposed Algorithm

We propose two algorithms for mining of interdimension association rules in transaction database .Those algorithms are :CombineDims, and GenFI.

First we apply the CombineDims algorithm to combine the selected dimensions in order to provide the framework for mining interdimension association rules. Then, we apply the GenFI algorithm to discover frequent itemsets in the transaction database. For the new information system, the searching of frequent itemsets is easy based on the concept of equivalence class.

4.1 CombineDims Algorithm

We prepare the data from the general table as follows:

1. Select the dimension d, d_1, d_2, ….. d_m From the general tables where ($d_1 = duser_1$) And ($d_2 = duser_2$) And…..($d_m = duser_m$) group by < $d_1, d_2 \ldots\ldots\ldots d_m$ > .This syntax create an initialized table IntTab for mining multidimensional association rule. Now we apply one distinct mapping code which is stored on MapTab for selected dimension as follows.
(age dimension/sex dimension/income dimension, buys(d),and mapping code)

('29/'M'/30k,'Laptop','0001')

Here we combine three dimensions: age, sex and income into one mapping code '0001'.The following are the details of our proposed algorithms. Note that notations in table 3 are used for our proposed algorithms.

Table 3. Notation

Notation	*Meaning*
D	*Sets of dimensions and its values* $\{d, d_1, d_2, \ldots d_m\}$
ComDim	*Combine Dimensions and its values*$\{d_1, d_2, \ldots d_m\}$
IntTab	*Initialize Table* $\{D, count\}$
MdTab	*Md Table* $\{d, MapCode\}$
KeyTab	*Key Table* $\{d\}$
MdTabProcess	*Process Md; contains*{ d, *List of MapCode}*
TmpLargeTab	*Temp Large Itemset Table {List of ComDim,Level,Sup}*

1.Procedure CombineDims
2. X={Total rows of table IntTab}
3.For I=1 to X Loop //on table IntTab
4. If !CheckMapCode($d_1, d_2 \ldots d_m$) then
5. GenMapCode($d_1, d_2 \ldots d_m$) ;
6. End IF;
7. End Loop;
8.For J= 1 to X Loop// on table IntTab
9.S=FindMapCode($d_1, d_2 \ldots d_m$);
10. Insert MdTab(IntTab($d.key$) ,MapTab(MapCode))
11. End Loop;

After creating MdTab, we use that table in the GenFI algorithm to discover frequent itemset on interdimension mining association rules in transaction database.

4.2 GenFI Algorithm
1. X={total rows of MdTab};
2. Y={total rows of table keyTab};//key table{ d }
3. N={total attributes of selected $d_m.key$};
4. For I = 1 to X Loop // on table MdTab
5. IF !CheckKey(d) then
6. Insert keyTab(d);
7. End IF;
8. End Loop;
9. For J = 1 to Y Loop// on KeyTab
10 Insert into $ListMapCode_j$
11. Select $MapCode_1, MapCode_2, \ldots, MapCode_m$
12. From MdTab a, KeyTab b
13. Where a.(d) = b.(d)
14. And b.(d) = $kaytab_j.(d)$;

15. Insert MdTabProces($kaytab_j.(d)$. $ListMapCode_j$);
16. EndLoop;
17. FI_Gen(MdTabProcess,TmpLargeTab(List of ComDim,Level,Sup),MinSup);
18. Transform_MapCode(TmpLargeTab,MapTab,LargeTab);

In FI_Gen candidate itemsets are generated by equivalence classes[10] and the searching method for candidate itemsets is similar toApriori algorithm.

After discovering all the large itemsets in the table LargeTab,we will have our interdimension association rule template as follows:

$$d_1(val), d_2(val), ..., d_m(val) \rightarrow d(val)$$

4.3 Mining of Association rules

The mining of association rules is usually a two phase's process. The first phase is for frequent itemsets generation. The second phase generates the rules using another user defined parameter ***minconf,*** which again affects the generation of rules. The second phase is easier and the overall performance of mining association rules is determined mainly by the first step [1].

5. Experimental Result

To evaluate the efficiency of the proposed method, the RSMAR, along with the Apriori algorithm, is implemented at the same condition. We use a sample sales database which contains three dimensions (i.e. customer dimension, product, dimension, Promotions dimension) and one sales fact table (see table 4).We perform our experiments using a Pentium IV 1,8 Gigahertz CPU with 512MB.

Table 4. Sales Database

Table Name	***Records***
Customer Dimension	***100***
Product Dimension	***50***
Promotions Dimension	***50***
Sales Fact Table	***1000***

The minimum support of Apriori algorithm is 0.45%, and the computation times and the numbers of frequent itemsets found by the two algorithms are shown in Figure 2.

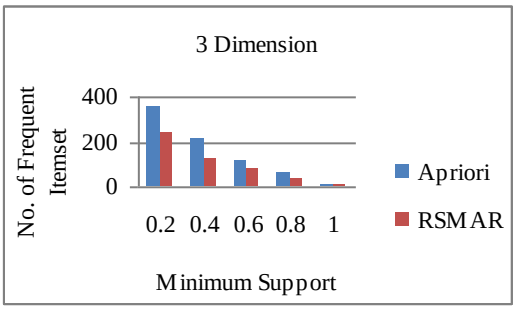

Figure 2 (a) No of frequent itemset.

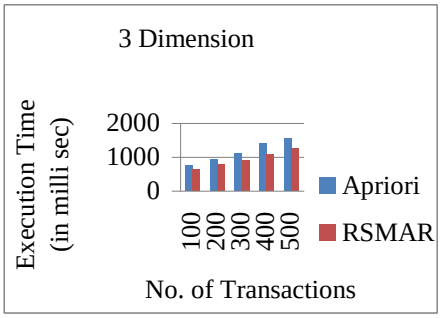

Figure 2 (b)Computation Time

Figure 2.(a) No of frequent itemset. (b)Computation Time

The experimental results in Figure 2 show that the RSMAR performs batter and more rapid than the Apriori algorithm. The RSMAR is not only eliminating considerable amounts of data, but also decreasing the numbers of database scanning, thus reducing the computation quantities to perform data contrasts and also memory requirements.

6. Conclusion

In this paper, the RSMAR is proposed to mining of interdimension association rules. Mining rules with the RSMAR algorithm is two step processes: First we apply the CombineDims *algorithm* to combine the selected dimensions in order to provide the framework for mining interdimension association rules. Then, we apply the GenFI algorithm to discover frequent itemsets in the transaction database. For the new information system, the searching of frequent itemsets is easy based on the concept of equivalence class. The algorithm provides better performance improvements. The gap between the RSASM and Apriori algorithms becomes evident with the number and size of pattems identified and the searching time reduced. In this paper, we still restricted our proposed

extended method to generate interdimension association rules. In future we will discuss and propose a method to generate hybrid-dimension association rules by assuming that hybrid-dimension association rules is hybridization between intradimension and interdimension association rules.

References

[1] Agrawal, R., Imielinski, T., Swami, A., "Mining Association Rules between Sets of Items in Large Databases", *SIGMOD'93*, pp. 207-216, 1993.

[2] Bodon, F., "A Fast Apriori Implementation", *FIMI'03*, November 2003.

[3] Han, Jiawei, Micheline Kamber, *Data Mining:Concepts and Techniques*, The Morgan Kaufmann Series, 2001

[4] 11. Agrawal, Rakesh, Ramakrishnan Srikant, *Fast Algorithms for Mining Association Rules in Large Databases*, Proccedings of 20th International Conference Very Large Databases, Morgan Kaufman, 1994, pp. 487-499.

[5] Codd, Edgar F., Communication of the ACM 13 (6), 1970, pp. 377-387.

[6] Intan, Rolly, A Proposal of Fuzzy Multidimensional Association Rules, *Jurnal Informatika*, Vol. 7 No. 2 (Terakreditasi SK DIKTI No. 56/DIKTI/ Kep/2005), November 2006.

[7] Jurgens, M. and Lenz, H.-J. (2001). Tree Based Indexes Versus Bitmap Indexes: A Performance Study. *International Journal of Cooperative Information Systems*, 10, 355–376.

[8] Pawlak, Z. (1982). Rough Sets. *Int. J. Computer and Information Sci*, 11, 341–356.

[9]] Munakata, T. (1998). Rough Sets. In: *Fundamentals of the New Artificial Intelligence*, 140–182. New York: Springer-Verlag.

[10] Xin Ma" Rough Set Model for]Discovering Single-dimensional and Multidimensional Association Rules*" 2004 IEEE International Conference on Systems, Man and Cybernetics

[11] Daniel Delic, Hans-J. Lenz, and Mattis Neiling Free" Improving the Quality of Association Rule Mining by Means of Rough Sets"

Anjana Pa ndey, born on Dec'18, 1978. She completed her Master in Computer Application. Her special fields of interest included Data mining. Presently working in Department of Information Technology at UITRGTU, Bhopal. Presently she is persuing PhD. In MCA from MANIT, Bhopal.

K..R.Pardasani was born on 13th September 1960 at Mathure, India. He completed his graduation, post graduation and PhD (mathematics) from Jiwaji university gwaliar India. his employment experience includes Jiwaji university Gwaliar MDI, Gurgaon and MANIT Bhopal India. Presently he is professor & Head of mathematics at MANIT, Bhopal and his current interest are data mining and computational biology.