2011 International Conference on Information and Network Technology
IPCSIT vol.4 (2011) © (2011) IACSIT Press, Singapore

Binary Association Rule Mining Using Bayesian Network

Venkateswara Rao Vedula [1] and Satish Thatavarti [2]

[1]Associate Professor Dept of Computer Science and Engineering, Sri Vasavi Engineering College, Pedatadepalli, Tadepalligudem—534101, Mobile No: +91-9866958546, Email: venkatvedula@yahoo.com

[2] Final M.Tech Student Dept of Computer Science and Engineering, Sri Vasavi Engineering College, Pedatadepalli, Tadepalligudem—534101, Mobile No: +91-9966145457, Email: satish5813@gmail.com

Abstract: The Trivial association rule mining which should be fixed in order to avoid both that early trivial rules is retained and also that interesting rules are not discarded. In fact the situations which use the relative comparison to express association rules are more complete than the situations that generate association rules that use the absolute comparison. The traditional concept of association rule mining will lose some information in generating association rules. The user does have to determine the degree of support and confidence thresholds before generating association rules. In our paper, we proposed new approach in finding association rules. This new approach uses the concept of rough set theory and Bayesian network classification to generate association rules. This provides a way for decision maker to get more information to generate association rules than traditional approach. The new approach for revving association rules has the ability to handle the certainty in the classifying process so that we can reduceinformation loss and enhance the result of data mining. This new algorithm can simulate the value of probability which is based on continuous data set.

Keywords: Data mining, Association rule mining, frequent item set, Electronic commerce, Rough set theory, Bayesian network, Bayesian classification.

1. Introduction:

Association rules of events/nodes can be regarded as probability rules due to their co-occurrence [1]. The real life example is the database of sales transactions. In such case, the objective is to find the items that are bought together. Such information is helpful in the development of marketing strategies with great success. 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 *intra dimension 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 intra dimension 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.

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”). In association rules, events are taken as variables but ifa complete rule is taken as a variable/node which is basedon probability rule with maximized probability to developthe relation among relation for further generalization inthe market strategy.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 multidimensional association rules will be discussed later in our next paper. The inter dimension 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.

This paper is organized as follows: Section 2 completely describes System Design and Implementation. Section 3 describes Proposed work(Association Rule Mining for generating Binary Association Rules). Section 4 presents experimental results and performance analysis. Section 5 presents conclusion and future scope.

2. System Design and Implementation:

The Following Figure shows Architecture of the Proposed System.

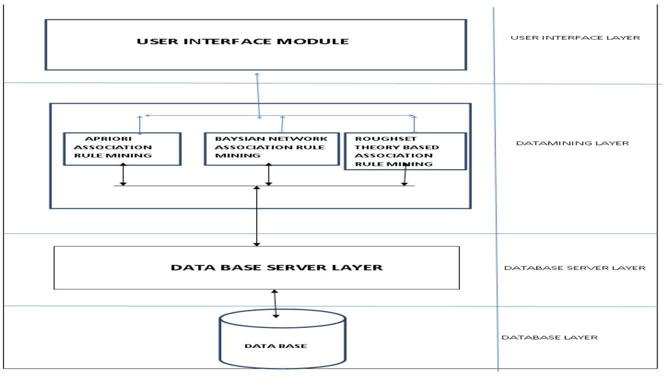


Figure 2.1 System Architecture

The system is composed of five layers

1) User interface layer 2) Data mining algorithms implementation layer 3)Data base server layer
4) Data base layer

User Interface Layer:

This layer is responsible for interaction with user and various calls to different graphical and visualization utilities. This module provides an interface for user to invoke with system and to execute queries based Association Rule Mining algorithms. The following are different services that this module offers.

1) To design an interface for Apriori algorithm. 2) To design an interface for Bayesian algorithm.
3) To design an interface for Rough set based algorithm

Data mining algorithms implementation layer: This layer is main layer that connects all system components together. This layer is divided into 3 sub modules. 1) Apriori Association Rule Mining Module 2) Bayesian Association Rule Mining Module 3) Rough Based Association Rule Mining Module

Database server layer : This layer uses a database server for maintaining transactional data items for use in Association Rule mining algorithm.

<u>Database layer :</u> This layer is collection of datasets that maintains transactional data items. We can also datasets for maintaining transactional data items.

3. Proposedwork:

3.1 Association Rule Mining Using Rough Set Theory : In 1982 Z.Pawalak 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

3.1.1 Model: The rough set method operates on data matrics, so called "Information System.It contains data about the universe of interest, condition attributes and decision attributes. 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= $\{x_1.x_2\ldots..x_n\}$, *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∪D where C= c1 ∧ c2 ∧ c3 ….. ∧cn and D= d1∧ d2 ∧ d3 …… ∧ dn.

V = $U v_p$ and v_p is a domain of attribute p.
P ε A

The function *f* performs a mapping code of conditionattributes such that *c*1, *c*2…*cn* into one simple attribute *C* which can be added directly into the informationsystem as one certain attribute, it will only posses onecolumn in the information system, analogous an item. f: U X at → V is a total function such that $f(x_{i,q})$ε V. A prerequisite for rule generation is a partitioning of in a finite number of blocks, so called equivalence classes, of same attribute values by applying an equivalence relation. We propose an algorithm for mining of interdimension association rules in transaction database i.eCombineDims

3.1.2 Combine Dims Algorithm (RSMAR):

We prepare the data from the general table as follows:

1. Select the dimension d, d1,d2,d3,…,dm From the general tables where d1=(duser1) And d2=(duser2) And…..dm=(duserm) group by <d1,d2,……dm>.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 algorithm.

```
1. Procedure CombineDims
2. X= {Total rows of table IntTab}
3. For I=1 to X Loop //on table IntTab
4. If! CheckMapCode (d1, d 2.... d m) then
5. GenMapCode (d1, d 2....dm);
6. End IF;
7. End Loop;
8. For J= 1 to X Loop// on table IntTab
9. S=FindMapCode (d1, d 2.... 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 item set on interdimension mining association rules in transaction database.

3.1.3 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.

3.2 Association Rule Mining Using Bayesian Network: Bayesian Network (BN) is used to build a modelusing a probability distribution over a set of variables. The benefit of BN is the compact representation of complex problem domains. Also, BN provides decision making, smooth, consistent and flexible applicability in the complex domains. On the other hand, association rules are based on antecedent and consequent part which have condition attributes and decision attributes respectively. Each association rule has a confidence percentage value which shows the rule effectiveness in terms of probability. Thus, association rules and Bayesian Network can be linked up due to probabilistic approach. Therefore, we focus on association rules to develop an automated data mining technique with the use of Bayesian Network. For this purpose, we suggested Associated Rules Binary Symmetric Matrix using K2 (ARBSM-K2) technique in order to generate Hierarchical Association Rules (HAR). This structure shows the relations among the association ruleswhich show more certainty in the hierarchy after maximizing the probability for each association rule treated as a node. Hence, this hierarchy may open new research dimensions to academia community by refining the existing techniques with the latest techniques to further obtain knowledge in any domain. For generating Hierarchical Association Rules first association rules as shown in Table 1 are converted into the binary dataset using their confidence% values in a text file as a dataset. Therefore, such conversion in Binary Dataset is shown in the following matrix where Δi j=Δji;

Table 3.2.1 Association Rules

Rule #	Conf. %	Antecedent (a)	Consequent (c)
1	60.61	RefBks=>	ChildBks, CookBks
2	90.91	ChildBks, RefBks=>	CookBks
3	57.14	ChildBks, CookBks=>	GeogBks
4	51.85	GeogBks=>	ChildBks, CookBks
5	88.46	CookBks, YouthBks=>	ChildBks
6	86.96	DoItYBks, GeogBks=>	CookBks
7	57.14	ChildBks, CookBks=>	DoItYBks
8	50	ArtBks=>	ChildBks, CookBks

Matrix

$$\begin{bmatrix} \Delta 11\% & \Delta 12\% & \Delta 13\% & \cdots & \Delta 1n\% \\ \Delta 21\% & \Delta 22\% & \Delta 23\% & \cdots & \Delta 2n\% \\ \Delta 31\% & \Delta 32\% & \Delta 33\% & \cdots & \Delta 3n\% \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Delta n1\% & \Delta n2\% & \Delta n3\% & \cdots & \Delta nn\% \end{bmatrix}$$

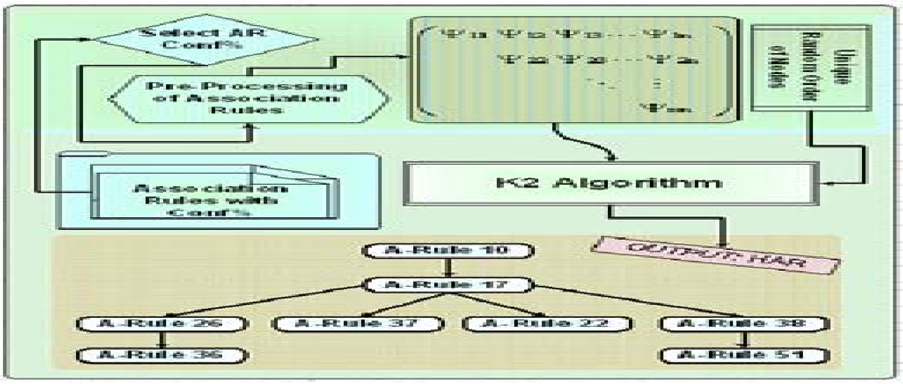
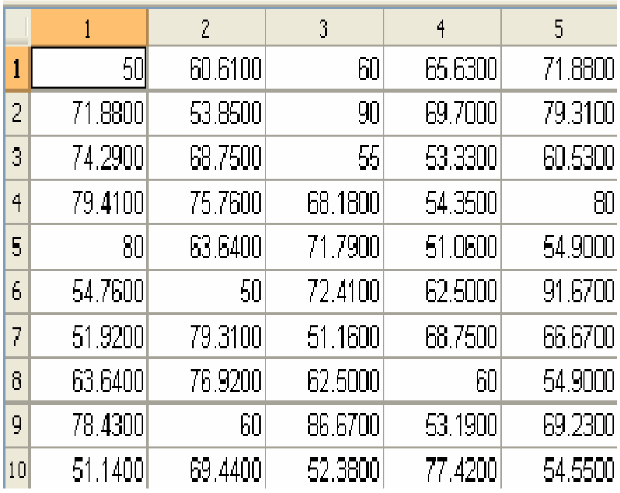


Figure 3.2.1 Proposed Model for HAR

4. Experimental Results

In order to implement the required Bayesian Network algorithm (HAR), we selected to use c# language in .Net environment. For this purpose, a text matrix file of association rules on confidence values in percentages is used. The following are the results of Bayesian Network algorithm implementation.



	1	2	3	4	5
1	50	60.6100	60	65.6300	71.8800
2	71.8800	53.8500	90	69.7000	79.3100
3	74.2900	68.7500	55	53.3300	60.5300
4	79.4100	75.7600	68.1800	54.3500	80
5	80	63.6400	71.7900	51.0600	54.9000
6	54.7600	50	72.4100	62.5000	91.6700
7	51.9200	79.3100	51.1600	68.7500	66.6700
8	63.6400	76.9200	62.5000	60	54.9000
9	78.4300	60	86.6700	53.1900	69.2300
10	51.1400	69.4400	52.3800	77.4200	54.5500

Figure4.1 Dataset used.

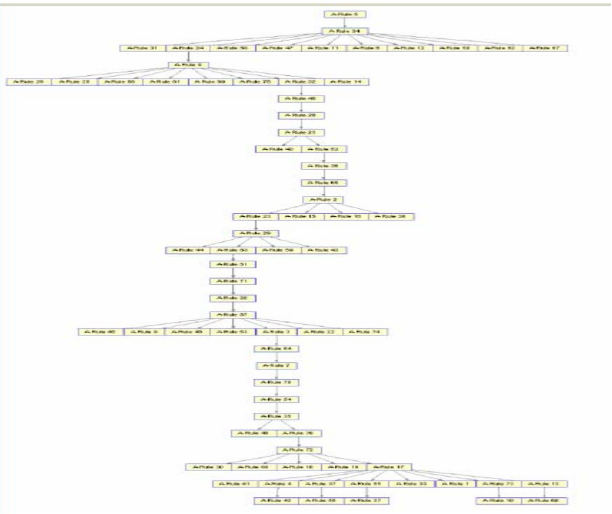


Figure4.2 Hierarchical Association Rules.

Figure 4.1 shows Data set used in the system and Figure 4.2 shows The generated Hierarchical Association Rules.

The following are the results of Rough set theory based association rule mining algorithm (Combined Dims algorithm);

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

Table4.1 Sales Database.

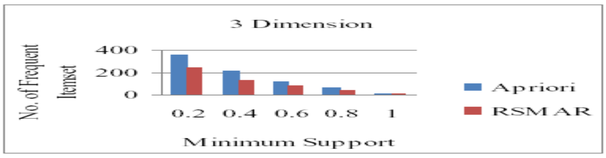


Figure 4.3(a) No of frequent item set.

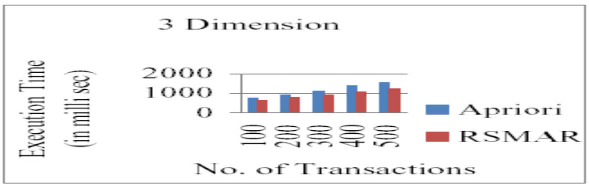


Figure 4.3(b) Computation Time

The following Figures (Figure 4.4 and Figure 4.5) show the comparison of performance of Apriori and Baysian Network Algorithms by using time complexity.

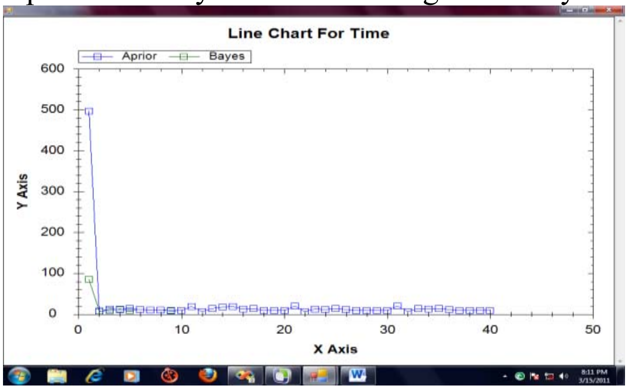


Figure 4.4

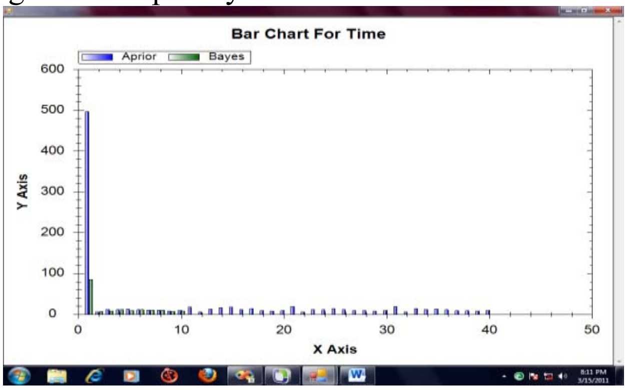


Figure 4.5

The following Figures (Figure 4.4 and Figure 4.5) show the comparison of performance of Apriori and Baysian Network Algorithms by using Space complexity.

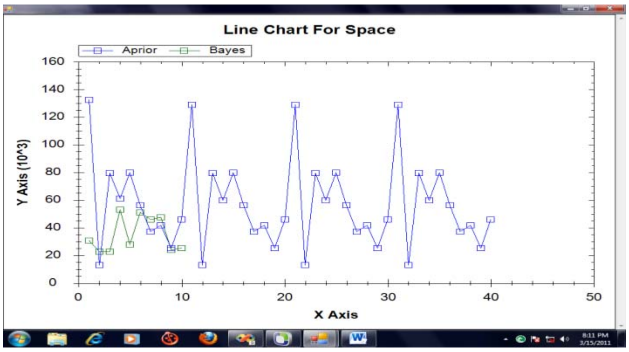


Figure 4.6

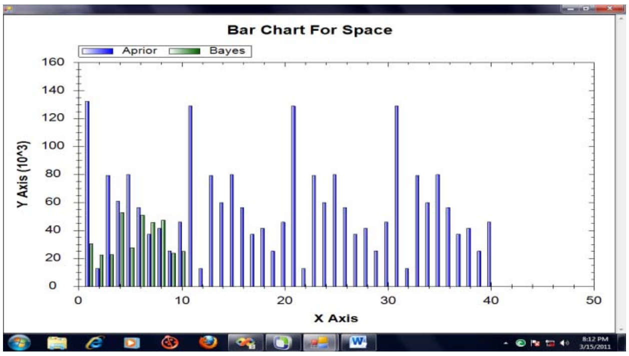


Figure 4.7

5. Conclusions and Future Work

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. 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. Also in this paper, we presented an automated HAR data mining technique (Baysian Network based Association Rule Mining algorithm). This technique is useful for obtaining further knowledge from association rules in a hierarchical way. Additionally, this technique shows the relation in uncertain environment among nodes (association rules) with the help of probability.

6. References

[1]. Lian, Wang; Cheung, David W.; Yiu, S.M.," An efficient algorithm forfinding dense regions for mining quantitative association rules",Computers and Mathematics with Applications, Vol.50, No.3-4, 2005,pp. 471-490.

[2]. Chen, Yen-Liang; Weng, Cheng-Hsiung, "Mining association rules fromimprecise ordinal data", Fuzzy Sets and Systems, Vol. 159, 2008, pp.460-474.

[3]. Boris Rozenberg, Ehud Gudes, "Association rules mining in verticallypartitioned database", Data & Knowledge Engineering, Vol.59, 2006, pp.378-396.

[4]. S. H. Liao, C. M. Chen., and .C. H. Wu, (2008) _ Mining customerknowledge for product line and brand extension in retailing , Expert systems with Applications, Vol. 35, Issue. 3. pp. 1763-1776.

[5]. De Cock, M.; Cornelis, C.; Kerre, E.E.,(2005)" Elicitation of fuzzyassociation rules from positive and negative examples" Fuzzy Sets andSystems, Vol. 149, Issue. 1. pp. 73-85.