Volume 2, Issue 1, January 2014

## **International Journal of Research in Advent Technology**

Available Online at: <a href="http://www.ijrat.org">http://www.ijrat.org</a>

# IMPROVEMENT IN ASSOCIATION RULE MINING BY MULTILEVEL RELATIONSHIP ALGORITHM

Deepak A. Vidhate<sup>1</sup>, Dr. Parag Kulkarni<sup>2</sup> <sup>1</sup> Department of Information Technology <sup>1</sup> P.D.V.V.P. College of Engineering, Ahmednagar <sup>2</sup> EKLaT Research Lab Pune <sup>1</sup> Email - dvidhate@yahoo.com,

#### **ABSTARCT:**

Association Mining is the discovery of relations or correlations among an item set. An objective is to make rules from given multiple sources of customer database transaction. It needs increasingly deepening knowledge mining process for finding refined knowledge from data. Earlier work is on mining association rules at one level. Though mining association rules at various levels is necessary. Finding of interesting association relationship among large amount of data will helpful to decision building, marketing, & business managing.

Mining the Data is also known as Discovery of Knowledge in Databases. It is to get correlations, trends, patterns, anomalies from the databases which can help to build exact future decisions. However data mining is not the natural. No one can assure that the decision will lead to good quality results. It only helps experts to understand the data and show the way to good decisions.

For generating frequent item set we are using Apriori Algorithm in multiple levels so called Multilevel Relationship algorithm (MRA). MRA works in first two stages. In third stage of MRA uses Bayesian probability to find out the dependency & relationship among different shops, pattern of sales & generates the rule for learning. This paper gives detail idea about concepts of association mining, mathematical model development for Multilevel Relationship algorithm and Implementation & Result Analysis of MRA and performance comparison of MRA and Apriori algorithm.

Keywords: Apriori Algorithm, Association rule, Bayesian Probability, Data mining, Multilevel learning

#### 1. INTRODUCTION

Association rule mining concept has been applied to market domain and specific problem has been studied, the management of some aspects of a shopping mall, and an architecture that makes it possible to construct agents capable of adapting the association rules has been used.

Data mining refers to extracting knowledge from large quantity of data. Interesting association can be discovered among a large set of data items by association rule mining. The finding of interesting relationship among large amount of business transaction records can help in many business decisions making process, such as catalog plan, cross marketing and loss leader analysis[1].

Machine Learning deals with the design of programs that can learn rules from data, adapt to changes, and improve performance with experience. In addition to being one of the initial thoughts of Computer Science, Machine Learning has become vital as computers are expected to solve increasingly complex problems and become more integrated into daily lives. These include identifying faces in images, autonomous driving in the desert, finding relevant documents in a database, finding patterns in large volumes of scientific data, and adjusting internal parameters of systems to optimize performance. Alternatively methods that take labeled training data and then learn appropriate rules from the data seem to be the best approach to solve the problems. Moreover, it needs a system that can adapt to varying conditions which is user-friendly by adapting to needs of their individual users, and also can improve performance over time[2].

A shopping mall is a cluster of independent shops, planned and developed by one or several entities, with a common objective. The size, commercial mixture, common services and complementary activities developed are all in keeping with their surroundings. A shopping mall needs to be managed and, the management includes solving incidents or problems in a dynamic environment[3].

Available Online at: <u>http://www.ijrat.org</u>

As such, a shopping mall can be seen as a large dynamic problem, in which the management required depends on the variability of the products, clients, opinions. Aim is to develop an open system, capable of incorporating as many agents as necessary, agents that can provide useful services to the clients not only in this shopping centre, but also in any other environment such as the labor market, educational system, medical care, etc. However, previous work has been focused on mining association rules at a single concept level. There are applications, which need to get associations at multiple concept levels. The focus was on working on mathematical model development for multilevel association rule mining. Multilevel Apriori algorithm and bayesian probability estimation is not combined in any of the previous work. It is the novel move towards the mining association rule. Efficiency of original Apriori algorithm has been increased due to multilevel architecture.

#### 2. ASSOCIATION RULE

Market basket analysis is useful for retailers to plan which item to put on sale at reduced price. If customer tends to purchase shirt of Bombay ding and jeans of Levis together, then having a sale on jeans may encourage the sale of shirt as well as jeans. Buying patterns reflects which items are frequent associated or purchased together. These patterns are represented in the form of association rules. For example, customer who purchase shirt-Bombay ding also tends to buy jeans Levis at the same time is represented in association rule (2.1) below.

Shirt-Bombay ding  $\Rightarrow$  jeans-levis

[supp=2%, conf=60%] (2.1)

Mining association rule is finding the interesting association or correlation relationship among large set of data items. Many industries are becoming interested in mining association rule from their database as massive amount of data constantly being collected & stored in database. Relationship among the business traction records can help to design catalog, loss leader analysis, cross marketing & other business decision making process. The discovery of such association can help retailers to develop marketing strategies by gaining insight into which items are frequently purchased together by customers. This information can increased sale by helping retailers to do selective marketing & plan their shelf space. One of the motivating examples for association rule mining is marker basket analysis[4].

Rule support & confidence are two measure rules. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for association rule means that 2% of all transactions under analysis show that shirt-Bombay ding and jeans-levis are purchased together. A confidence of 60% means that 60% of customers who purchased shirt-Bombay ding also bought jeans Levis. Typically, association rule are considered interesting if they satisfy both a minimum support threshold and a minimum assurance threshold. Such threshold can be located by users or area expert.

Let  $I = \{i1, i2, i3, \dots, i_d\}$  set of all items in dataset

 $T = \{t1, t2, t3..., t_n\}$  set of all transactions

Each transaction  $t_i$  contains a subset of items chosen from I. A transaction  $t_j$  is said to contain an itemset X if X is subset of  $t_j$ .

Association rule is an implication of the form of

 $X \Longrightarrow Y$ , where  $X \subseteq I$ ,  $Y \subseteq I \& X \cap Y = \Phi$ 

The rule  $X \Rightarrow Y$  holds in the transaction set T with support s, where s is percentage of transactions in T that contain X U Y. The rule  $X \Rightarrow Y$  has confidence c in the transaction set T if c is percentage in transactions in T containing X which also contain Y. i.e

Support $(X \Longrightarrow Y) = P(X \cup Y)$	(2.2)
Confidence $(X \Longrightarrow Y) = P(Y X)$	(2.3)
D1 4 4 6 1 4 1 1	1 1 11 / .

Rules that satisfy both minimum support threshold (min\_sup) and a minimum confidence threshold (min\_conf) are called strong.

Itemset is nothing but set of items. If it contains n item is a n-itemset. The set {shirt-Bombay ding, jeanslevis} is 2 itemset. The occurrence of itemset is the number of transactions that contain the itemset. This is known as frequency or support count of the item set. It satisfies lowest amount of support if the occurrences frequency of itemset is greater than or equal to the product of min\_sup & total no of transactions in T. If an itemset satisfy the minimum support then it is frequent itemset. Association mining has two steps process. In first step, find all frequent item sets. All of these item sets will arise at least as frequently as a pre-determined

Available Online at: <u>http://www.ijrat.org</u>

minimum support count. In second step, generate strong association rules from the frequent item sets and must satisfy lowest amount of support and minimum confidence. The overall performance of mining association rule is determined by the first step[5].

### 3. MULTILEVEL RELATIONSHIP ALGORITHM

To improve the mining of association rules new mining algorithm has been developed as Multilevel Relationship Algorithm which works in three stages. In first two stages it utilizes apriori algorithm for finding out frequent itemsets. Third stage of MRA uses bayesian probability to find out the dependency & relationship amongst different shops and generates the rules for learning.

Let the system S be represented as

 $S = \{I, O, f_s \mid \Phi_s\}$ 

I = Input Datasets

O = Output Patterns

 $O = f_s(I) \quad \forall \Phi_s$ 

 $f_s: I \rightarrow O$  be ONTO function

Objective was to find out pattern of sale from given dataset of three different shops for particular time period. Input dataset  $I = \{X,Y,Z\}$  such that  $X = \{x1,x2,x3\}$ ,  $Y = \{y1,y2,y3\}$  and  $Z = \{z1,z2,z3\}$ 

Success output  $O = \{P(X0|Y0), P(X0|Z0), P(X1|Y1), P(Y1|Z1), ..., \}$ 

Multilevel Relationship Algorithm is applied on given input dataset i.e.  $I=\{X,Y,Z\}$  where  $X = \{x1,x2,x3\}$ ,  $Y = \{y1,y2,y3\}$  and  $Z = \{z1,z2,z3\}$ .

First stage gives Level 1 association amongst items in the same shop using knowledge base. It is called as local frequent itemsets generated in first phase. During second stage it uses individual knowledge base and level 1 association that was generated in stage I from same shops to find out the frequent item sets i.e. x1(0), x2(3), x3(1).....etc. It is called as global frequent itemsets.

Stage 1:

At first stage it find out Level 1 association amongst items in the same shop i.e. internal relationship between the same item types i. e. x1(0,...,n), x2(0,...,n), x3(0,...,n) within the Cloth shop (X) i.e.  $O = f_s(X)$ . It find out internal relationship between the same item types i. e. y1(0,...,n), y2(0,...,n), y3(0,...,n) within the Jewelry shop (Y) i.e.  $O = f_s(Y)$ . Also it find out the internal relationship between the same item types i. e. z1(0,...,n), z2(0,...,n), z3(0,...,n) within the Footwear shop (Z) i.e.  $O = f_s(Z)$ .

Stage 2:

During second stage it uses individual knowledge base and level 1 association is generated in stage 1 of same shop to find out the frequent item sets i.e. x1(0), x2(3), x3(1).....etc is called as global frequent itemsets. It gives sets of frequent item sets for the Cloth shop for different items i.e. Fx as  $O = f_s(x1,x2,x3)$ . It gives sets of frequent item sets for the Jewelry shop for different items i.e. Fy as  $O = f_s(y1,y2,y3)$ . And also gives with sets of frequent item sets for the Footwear shop for different items i.e. Fz as  $O = f_s(z1,z2,z3)$ .

Stage 3:

It is necessary to determine dynamic behavior of  $F_i$  for particular season. External Dependencies amongst Items Xi Yi...... Xn Yn has been found with Bayesian probability. New patterns are generated by Bayesian probability through which learning rules are predicted & interpreted.

Available Online at: <a href="http://www.ijrat.org">http://www.ijrat.org</a>

Working of Multilevel Relationship Algorithm

Let the sale of *Item X* at Cloth shop affects sale of *item Y* at Jewelry shop and *item Z* at Footwear.

- 1. Apriori association mining algorithm is applied on each item in cloth shops separately i.e. Jean(X0), Tshirt(X1), Shirt(X2) and so on from the given large item sets. It was applied at two levels / phases in the same shop.
- 2. After applying Apriori algorithm at first level for different support value it provide with the internal dependency amongst individual items & generate the individual knowledge base i.e.  $x1(0) \rightarrow x1(1), x2(0) \rightarrow x2(1), x3(0) \rightarrow x3(1)$  .....etc. It is called as local frequent itemsets generated in first phase.
- 3. At second level Apriori algorithm was applied on newly generated individual knowledge base to find out the frequent item sets i.e. x1(0), x2(3), x3(1).....etc. It is called as global frequent itemsets.
- 4. It provided with sets of frequent item sets for the Cloth shop for different items i.e. Fx.
- 5. Similarly the algorithm is applied on Jewelry shop(Y) & Footwear shop(Z) to determine frequent itemset on different items.
- 6. First Level output of Apriori algorithm provided internal association amongst the items i. e.  $y_1(0) \rightarrow y_1(1), y_2(0) \rightarrow y_2(1), y_3(0) \rightarrow y_3(1) \& z_1(0) \rightarrow z_1(1), z_2(0) \rightarrow z_2(1), z_3(0) \rightarrow z_3(1)$ .....etc for Jewelry & Footwear shop respectively.
- 7. Second level input of Apriori algorithm provided from newly generated individual knowledge base, the frequent item sets i.e. y1(0), y2(3), y3(1), z1(1), z2(5)
- 8. It gives with sets of frequent item sets for the Jewelry & Footwear shop for different items i.e. Fy & Fz.
- 9. The context is generated under uncertainty in the form of frequent item sets Fx, Fy & Fz. System constraints applied here are sale of items in a day, week, month or any particular season. This context is refereed as F<sub>i</sub> which is not constant, i.e. it changed seasonably.
- 10. Hence it is necessary to determine dynamic behavior of F<sub>i</sub> for particular season.
- 11. External Dependencies amongst Items  $Xi \rightarrow Yi...Xn \rightarrow Yn$  is found with Bayesian probability.
- 12. New patterns are generated by Bayesian probability though which learning rules could be predicted & interpreted.

### 4. ARCHITECTURE OF MRA



Fig. 1 : MRA Architecture Diagram

Figure 1 shows the flow diagram which depicted the development of Multilevel Relationship Algorithm (MRA). Multilevel Relationship algorithm worked in three stages.

In first two stages it utilized association rule mining algorithm for finding out frequent itemsets. Datasets of three shops i.e. Cloth, Jewelry & Footwear were given as an input to the stage I and Level 1 association between individual items had been found out. Level 1 association between individual items was given as an input to stage II and frequent itemsets had been found out. These frequent itemsets had generated new sale context. In stage III it used bayesian probability to find out the external dependency & relationship amongst different shops, pattern of sale and generated the rules for cooperative learning. The algorithm consists of three sub modules: MRA Stage I, MRA stage II, Interdependency Module

Available Online at: <a href="http://www.ijrat.org">http://www.ijrat.org</a>

MRA Stage I:

At first stage it finds Level 1 association amongst items in the same shop i.e. Internal relationship between the same item types i. e.  $x1(0, \ldots, n), x2(0, \ldots, n), x3(0, \ldots, n)$  within the Cloth shop (X) i.e.  $O = f_s(X) = f_{stage\___algorithm\_apriori}(X)$   $\therefore O = f_{stage\___algorithm\_apriori}\{x1(\ldots, n)\} = \{x1(0) \rightarrow x1(1), x1(3) \rightarrow x1(2), \ldots\}$   $O = f_{stage\___algorithm\_apriori}\{x2(0, \ldots, n)\} = \{x2(2) \rightarrow x2(4), x2(2) \rightarrow x2(4), \ldots\}$   $O = f_{stage\___algorithm\_apriori}\{x3(0, \ldots, n)\} = \{x3(0) \rightarrow x3(3), x3(1) \rightarrow x3(5), \ldots\}$ MRA Level 1 finds internal relationship between the same item types i. e.  $Y1(0, \ldots, N), Y2(0, \ldots, N), Y3(0, \ldots, N)$  within the Jewellery shop (Y) i.e.  $O = f_s(Y) = f_{stage\_\__algorithm\_apriori}(Y)$   $\therefore O = f_{stage\_\__algorithm\_apriori}\{y1(0, n)\} = \{y1(1) \rightarrow y1(3), y1(2) \rightarrow y1(5)\}$   $O = f_{stage\_\_algorithm\_apriori}\{y2(0, n)\} = \{y2(0) \rightarrow y2(1), y2(3) \rightarrow y2(7)\}$  $O = f_{stage\_\_algorithm\_apriori}\{y3(0, n)\} = \{y3(2) \rightarrow y3(3), y3(1) \rightarrow y3(4)\}$ 

MRA Level 1 also finds internal relationship between the same item types i. e. z1(0....n), z2(0....n), z3(0....n) within the Footwear shop (Z)

 $\begin{array}{l} O = f_s(Z) = f_{stage\_lalgorithm\_apriori}(Z) \\ \therefore O = f_{stage\_Lalgorithm\_apriori}\{z1(0,\ldots,n)\} = \{z1(0) \rightarrow z1(2), z1(2) \rightarrow z1(4),\ldots\} \\ O = f_{stage\_Lalgorithm\_apriori}\{z2(0,\ldots,n)\} = \{z2(1) \rightarrow z2(4), z2(1) \rightarrow z2(3),\ldots\} \\ O = f_{stage\_lalgorithm\_apriori}\{z3(0,\ldots,n)\} = \{z3(0) \rightarrow z3(3), z3(2) \rightarrow z3(5),\ldots\} \end{array}$ 

#### MRA Stage II:

During second stage it uses individual knowledge base and level 1 association is generated in stage 1 of same shop to find out the frequent item sets i.e. x1(0), x2(3), x3(1).....etc is called as global frequent itemsets. It gives sets of frequent

item sets for the Cloth shop for different items i.e. Fx as below.  $O = f_{s}(x1,x2,x3)$   $O=f_{phase_II\_algorithm\_apriori}\{x1,x2,x3\}$   $=\{x1(0) \rightarrow x2(1),x2(3) \rightarrow x3(2), x3(0) \rightarrow x2(2), \dots,\}$ MRA Stage II gives sets of frequent item sets for the Jewelry shop for different items i.e. Fy as below  $O = f_{s}(y1,y2,y3)$   $O=f_{phase\_II\_algorithm\_apriori}\{y1,y2,y3\}$   $=\{y1(0) \rightarrow y2(1),y2(3) \rightarrow y3(2), y3(0) \rightarrow y2(2), \dots,\}$ MRA Stage II also gives with sets of frequent item sets for the Footwear shop for different items i.e. Fz as below  $O = f_{s}(z1,z2,z3)$   $O=f_{phase\_II\_algorithm\_apriori}\{z1,z2,z3\}$ 

MRA Stage 3:

Interdependency by Bayesian Probability

 $= \{z_1(0) \rightarrow z_2(1), z_2(3) \rightarrow z_3(2), z_3(0) \rightarrow z_2(2), \dots \}$ 

It is necessary to determine dynamic behavior of  $\mathbf{F}_i$  for particular season. External Dependencies amongst Items  $Xi \rightarrow Yi....Xn \rightarrow Yn$  is found with Bayesian probability. New patterns are generated by Bayesian probability through which learning rules are predicted & interpreted. Dependency between itemsets of Cloth shop (Fx) and Jewelry shop (Fy) is find out as

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$
$$P(x1, x2, \dots xn|y1, y2, \dots yn) = \frac{P(y1, y2...yn|x1, x2...xn)P(x1, x2...xn)}{P(y1, y2 \dots yn)}$$

### Volume 2, Issue 1, January 2014

# **International Journal of Research in Advent Technology**

Available Online at: http://www.ijrat.org

Bayesian probability finds out interdependency between itemsets of Jewelry shop (Fy) and Footwear shop (Fz) as

$$P(Y|Z) = \frac{P(Z|Y)P(Y)}{P(Z)}$$

 $P(y1, y2, ..., yn|z1, z2, ..., zn) = \frac{P(z1, z2, ..., zn)}{P(z1, z2, ..., zn)P(y1, y2, ..., yn)}$ 

Bayesian probability also finds out interdependency between itemsets of Footwear shop (Fz) and Cloth shop (Fx) as

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)}$$

$$P(z1, z2, ..., zn|x1, x2, ..., xn) = \frac{P(x1, x2..., xn|z1, z2..., zn)P(z1, z2..., zn)}{P(x1, x2, ..., xn)}$$

#### 5. EXPERIMENTAL RESULTS

The experimental results that have been obtained through implementing MRA and Apriori algorithm are presented in this section. Multilevel relationship algorithm applied for finding the frequent itemset and external dependency amongst them. It comes up with pattern which can be further useful for leaning in cooperative system. The results obtained for strength, support and interdependency of itemsets for both the algorithms. Performance of Apriori and MRA has compared for these factors i.e. strength, support and interdependency. *Dataset Organization* 

Association mining data is generally obtained from databases created for other uses and manipulate into a suitable representation through pre processing techniques. The resulting dataset is expressed as items that they contain. Experiments have been conducted datasets of Cloth, Jewellery shops. Each data sets have the five attributes i.e. Transaction ID, Item, Brand, Quantity and date of purchase. Dataset contains various items with different brand purchased with diverse quantity during the specified period by the customers. Snap shots of each data set is given in following table.

Table 5.1: Cloth Shop (X)							
Transaction ID	Item	Brand		Quantity	Date		
1	Shirt	Bombay Dying		3	16/10/2012		
1	Jeans	Denis		1	16/10/2012		
2	Jeans	Peter England		2	17/10/2012		
2	Tshirt	Pepe Jeans		1	17/10/2012		
3	Shirt	Pan America		2	18/10/2012		
3	Tshirt	Being Human		1	18/10/2012		
4	Jeans	Levis		2	19/10/2012		
Table 5.2: Jewellery Shop (Y)							
Transaction ID	Item		Brand	Quantity	Date		
1	Bracelet		Nakshtra	3	16/10/2012		
1	Ear Rings		Gitanjali	2	16/10/2012		
2	Pendant		Asmi	2	17/10/2012		
2	Bracelet		Gilli	2	17/10/2012		
3	Diamond Ring		TBZ	1	18/10/2012		
3	Ear Rings		Asmi	2	18/10/2012		
4	Ear Rings		Nakshtra	2	19/10/2012		

Available Online at: <u>http://www.ijrat.org</u>

Following Graphs show the result comparison between Apriori and MRA.



Fig. 2: Comparison of Apriori & MRA in terms of item strength with support for cloth shop



Fig. 3: Comparison of Apriori & MRA in terms of item strength with support for jewellery shop



Fig. 4: Comparison of Apriori & MRA in terms of time (ms) & support for cloth shop

Available Online at: <u>http://www.ijrat.org</u>



Fig. 5: Comparison of Apriori & MRA in terms of time (ms) & support for jewellery shop

The experiment results show that MRA performs better than the Apriori algorithm towards improvement of mining association rule. Fig. 2 and 3shows comparison of Apriori & MRA in terms of item strength with support for cloth and jewellery shop. Item relative strength for minimum support count of MRA is always greater than Apriori algorithm. Increase in minimum support count decreases the item relative strength for both MRA and Apriori algorithms. Fig. 4 and 5 shows comparison of Apriori & MRA in terms of time (ms) & support for cloth and jewellery shop. Time required by MRA is always less than the time required by Apriori algorithm. As number of support increases it decreases the time requirement for both MRA and Apriori algorithms.

#### CONCLUSION

In this paper we proposed an efficient new Multilevel Relationship Algorithm. This is new approach applied to the set of data from different shops for finding frequent item sets and finding external dependencies amongst them. It comes up with patters which can be further useful for learning in cooperative algorithms. The classical apriori algorithm widely used for association rule mining. Though this algorithm is good to find the frequent item sets with minimum support it does not provide with dependencies between different frequent itemsets. The main contribution of this paper is that Multilevel Apriori algorithm and Bayesian probability estimation has not combined in any of the previous work.

#### REFERENCES

- [1] Aaron Ceglar & John F. Roddick "Association Mining" in ACM Computing Surveys, Vol. 38, No. 2, Article 5, Publication date: July 2006.
- [2] Baoqing Jiang, WeiWang and Yang Xu "The Math Background of Apriori Algorithm"
- [3] Mining Frequent Patterns without Candidate Generation Jiawei Han, Jian Pei, Yiwen Yin
- [4] R. Agrawal and R. Srikant "Fast algorithms for mining association rules in large databases" In Proceedings of the Twentieth International Conference on Very Large Databases, pages 487–499, Santiago, Chile, 1994.
- [5] R. Agrawal, T. Imielinski, and A. Swami "Mining associations between sets of items in massive databases" In Proc. of the ACM SIGMOD Int'l Conference on Management of Data, 1993.
- [6] Rakesh Agrawal, Christos Faloutsos, "Efficient similarity search in sequence databases" In Proc. Of Fourth International Conference on Foundations of Data Organization & Algorithms, Chicago, October 1993.
- [7] Rakesh Agrawal, SaktiGhosh, Tomasz Imielinski, BalaIyer, and Arun Swami "An interval classifer for database mining applications" Proc. of the VLDB Conference, pages 560-573, Vancouver, Canada, August 1992.
- [8] Rakesh Agrawal, Tomasz Imielinski and Arun Swami "Database mining: A performance perspective" published in IEEE Transactions on Knowledge and Data Engineering, 5(6):914 925, December 1993. Special Issue on Learning and Discovery in Knowledge-Based Databases.