JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 550 - 556
https://publishoa.com
ISSN: 1309-3452

Association Rule Generation using Pattern Mining Apriori Technique

Amit Verma[a], Raman Kumar[b]

[ab]Department of Computer Science and Engineering, IKGPTU, Kapurthala, Punjab, India
Email Addresses: verma0152@gmail.com, er.ramankumar@aol.in

ABSTRACT

The Apriori algorithms is a most powerful and influential algorithm used in mining of association rules. This works on the concept of identifying the frequent datasets from some transactional database. Some association rules are derived from these frequent sets, these derived rules must satisfy some criteria like minimum support value and confidence threshold etc. [1, 2]. The large batches of data need to be refined and filter to make use of it for some specific purposes. These purposes can be helpful in decision making processes of business environments. Useful patterns can be searched in large data batches which can further be used in business to learn more about the customers so that more useful and effective marketing strategies can be developed, cost can be reduced and sales can be increased. In this paper, we will study the analytical approach generally called frequent pattern mining, previously known as 'association' and explain study of how association rules are generated using the concept of Apriori Algorithm.

Keywords: frequent patterns, support, confidence, Apriori algorithm, association rule.

INTRODUCTION

Data Mining has a very good method of extracting frequent patterns from the huge databases [2]. These frequent patterns are basically frequent itemsets which have high frequency in large databases. Frequent pattern mining now can be defined as data mining technique which aims in developing or extracting frequent itemsets from a data base [3, 4]. These frequent itemsets plays an important role various data Mining jobs and are in relation with interesting patterns in the data such as association rules.

The technique of frequent pattern mining is concerned with the item sets are found in various types of databases like relational, transactional and data repositories, it extracts frequent patterns from such sources [5, 6]. Provided a set of transactions this technique focus to develop some association rules based on presence of a particular item based on presence of some other item [7].

Concept of frequent pattern mining can be clearer with an example. Whenever we visit to some grocery shop, we generally have a list of out required items to purchase. On the other side grocery owner has some other list depending on preference of the visiting customers. A customer should buy healthy items for the food whereas a bachelor would like to purchase beer and chips etc. The problem is to understand this buying pattern of the customers which can increase the sales rate of the business by various methods. Suppose item X and item Y are member of a single purchase transaction.

- Both the items X and Y can be placed in a row to promote the sale of one item with another.
- Various offers or discounts can apply to one out of two products.
- Some promotional advertisement can be done like in catalogue design both products (X and Y) can present together with offers.
- Both products X and Y can be combined in development of new product like X in flavour of Y.

Now after understanding the frequent patterns the problem is to how do we uncover associations between the products [8].

Association Rule Generation

The Analysis for Association rule [9] is a technique to discover how items are associated within the transaction. Following are the three ways to find association among items.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 550 - 556
https://publishoa.com
ISSN: 1309-3452

(i) Support: Let us for example item X and item Y appears frequently together in the transactional database. This frequency can be quantified using the metric support [10, 21].
Support of an itemset can easily be calculated by dividing transactions containing itemset with total transactions.

$$\textbf{Support } (\mathbf{X \rightarrow Y}): = \frac{\text{Transaction with X and Y}}{\text{Total Transactions}}$$

(ii) Confidence: This calculates how likely both items X and Y purchased together. Transactions having both items (X and Y) are divided with transactions having item X only [10].

$$\textbf{Confidence } (\mathbf{X \rightarrow Y}) = \frac{\text{Support of XUY}}{\text{Support of X}}$$

One of the drawback of confidence is that it sometimes only present the popularity of item X but not of item Y. Accounting the popularity of both items we use another measure also known as lift.

(iii) Lift: This measure controls the popularity of items when purchased together. It basically indicate how likely one item (Y) is purchased with another item (X) and controls popularity of first item.

$$\textbf{Lift } (\mathbf{X \rightarrow Y}): \quad = \frac{\text{Support of items (X, Y)}}{\text{Support X} * \text{Support Y}}$$

The calculated lift value states how likely items are purchased together.

- Lift value 0 → indicates no association between items.
- Lift value > 1 → Y more likely with X.
- Lift value < 1 → Y is not likely with X.

Let us explain these terms using a transactional database, suppose we have a transaction database of few products like milk, bread, butter and beer. Table below show the few transactions of these items from a database.

Transaction ID	Milk	Bread	Butter	Beer
1	Y	N	N	N
2	N	Y	N	N
3	N	Y	N	N
4	Y	Y	Y	Y
5	Y	Y	Y	N
6	Y	Y	Y	N

Table 1: Transaction Database

In the table above a transaction database is shown containing six transactions of items milk, bread, butter and beer.
Support(X): Transactions with (X) \ Total transactions.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 550 - 556
https://publishoa.com
ISSN: 1309-3452

Item	Support
Milk	0.6
Bread	0.8
Butter	0.5
Beer	0.3

Table 2: Support of Individual Item.

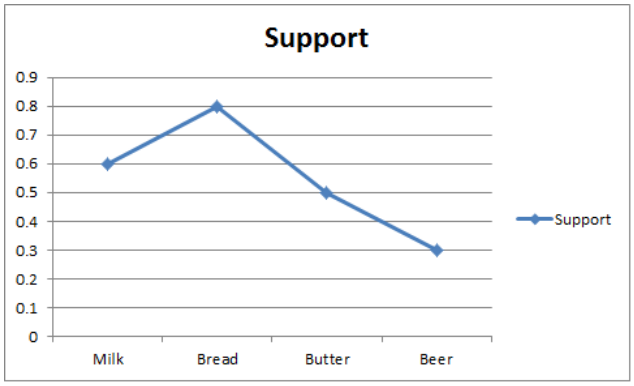

Fig.1 Calculated Support of Items.

Support (Milk → Bread) is 0.5 means that bread and butter are purchased simultaneously 50% of all the transactions. Three transactions out of six have entry Y for bread and butter.
Confidence (Milk → Bread) is 0.75 means that if 100 transactions contain the item Milk then out of these 75 transactions also contains the item Bread.

Transaction	Support	Confidence	lift
Milk → Bread	0.5	0.75	1.04
Milk → Butter	0.3	0.5	1.66
Milk → Beer	0.16	0.25	0.88

Table 3: Three Association Rule.

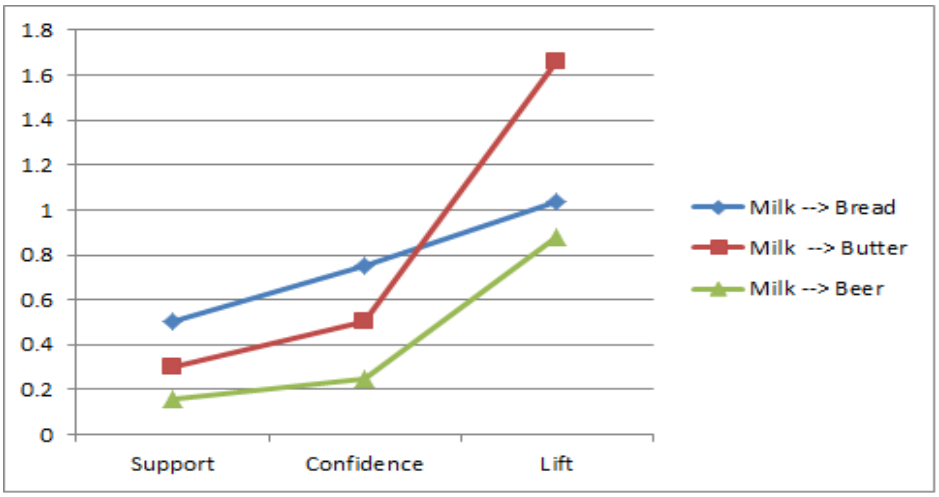

Fig. 2 Association between Items

Table 3 shows the three association rules [11, 12] with their support, confidence and lift values calculated. The value calculated of measure lift defines the likeliness of items as in our fig. 2 shows that Milk → Bread has lift value greater than 1 which means that customer like to buy bread when he purchases milk. If this value is below 1 then one item is

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 550 - 556
https://publishoa.com
ISSN: 1309-3452

unlikely purchased with another e.g. Milk → Beer in example. Lift value equals to 1 implies that there is no association between the items.
Here from table 3 it is clear that Milk is having good association with bread and butter and their sales can be promoted by placing the items together or some other sales promotion offers can be applied as customer who purchases milk likely to purchase bread and butter also.
Association rule mining is a two-step process which says that first discover all the frequents itemsets and then develop an association rule with discovered itemsets.

Apriori Algorithm: A frequent pattern mining technique, Apriori is the first frequent pattern mining algorithm discovered by R Agrawal and R Srikant which works in two phases known as join and prune [12]. It used iterative approach and reduces the search space for discovering the frequent pattern itemsets. Here are few steps which explain the working of algorithm:
Join Phase: Suppose k itemsets are given in the transaction database, then output of this step will be k+1 itemsets after joining each item with itself.
Prune Phase: A minimum support threshold is defined here and the items are scanned for their count in the database. If any item is found with less value then min support value it is called infrequent and not included in the itemset. This step removes the infrequent itemsets and minimizes the search space.
Detailed steps explaining Apriori method
Step 1) During first iteration each itemset is consider to have one item and algorithm count the occurrence of each item in the database.
Step 2) Define a minimum support value here and itemsets having greater and equal count to minimum support value are selected for next iteration. Itemsets with less count are discarded or pruned.
Step 3) Next itemsets with 2-items are discovered with frequent items. For this in join step 2-itemset is discovered by combining the items by itself.
Step 4) Here the 2-itemset sets are pruned comparing with minimum support value. Itemsets having support value less than the minimum support value are removed. Now the result contains only the frequent 2-itemsets.
Step 5) In the next iteration 3-itemset is formed with join and prune method. This step also applies the anti-monotone property which implies that if a subset is frequent then all its supersets are frequent otherwise discard the itemset and its supersets. While discovering 3-itemsets all its subsets of 2-itemsets must fall in minimum support criteria.
Step 6) Next step discovers the 4-itemsets using 3-itemsets and applying join and prune method similarly in step 5. The algorithm will stop when most frequent itemsets are discovered [13, 14] .
Explanation of the steps of Apriori with an example: let's assume the minimum support threshold value 50% and confidence threshold is 60%.

Transactions	List of Items
T1	i1, i2, i3, i5
T2	i2, i3, i4
T3	i4
T4	i1, i2, i4, i5
T5	i1, i2, i3
T6	i1, i2, i3, i4

Table 4: Transaction Database with itemset

Here already defined minimum support threshold value is 50% i.e., .5 * 6 = 3.
Step 1) Count every item of database is i1 = 4, count of i2 is 5, count of i3 is 4, count of i4 is 4 and count of i5 is 2.
Step 2) Now in the prune step we can discard the item i5 which is having less support value than the defined support value.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 550 - 556
https://publishoa.com
ISSN: 1309-3452

Step 3) Apply the join step and discover the 2-itemset.

2-itemsets	Count in Transaction Database
i1, i2	4
i1, i3	3
i1, i4	2
i2, i3	4
i2, i4	3
i3, i4	2

Table 6: 2-itemsets with count in data base.

Step 4) Apply pruning in above table and remove the itemsets which are less than the minimum support threshold value i.e. {i1, i4} and {i3, i4} have count 2 which is less than 3.
Step 5) From above table discovers the 3-itemset now their respective counts in the transaction database [15].

3-itemset	Count in transaction database.
i1, i2, i3	3
i1, i2, i4	2
i1, i3, i4	1
i2, i3, i4	2

Table 7: 3-itemset with count value.

Step 6) Now apply the pruning on above table and discover that only one 3-itemset i.e. {i1, i2, i3} is there having support value greater than the minimum support threshold all other itemsets are removed.
Now discover the association rule from the above discovered 3-itemset {i1, i2, i3}
Discovered association rule are (i1, i2) → (i3) having confidence 75%.
(Support (i1, i2, i3) / support (i1, i2) = (3/ 4)* 100 = 75%)

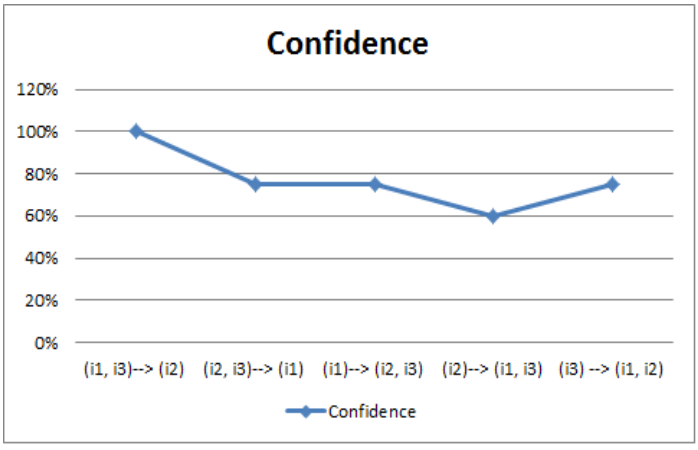

Fig. 3 Association rules with confidence.

From above figure it is clear that the association rules discovered meet the criteria of having confidence value more or equal to 60 %

- (i1, i3) → (i2) having confidence 100%
- (i2, i3) → (i1) having Confidence 75%.
- (i1) → (i2, i3) having confidence 75%.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 550 - 556
https://publishoa.com
ISSN: 1309-3452

- (i2) → (i1, i3) having Confidence 60%.
- (i3) => (i1, i2) having Confidence 75%.

HOW DOES PATTERN MINING SUPPORT BUSINESS ANALYSIS?

Pattern Mining is applied in business analysis processes and plays important role in making business decisions. This method is applied for data analysis of various business activities in organizations like

1. To analyse the items which are purchased together in a single cart or basket [16]. Association between items of a single cart is analysed for association e.g. between bread and butter purchased together in above transaction database this is known as **Basket Data Analysis.**

2. To analyse **cross marketing and selling** with other business organizations which also complement your own business [17]. For example, a motor spare part company also promote business of car painting goods as both have common final application.

3. Identification of association between items also helps in **designing catalogue** of companies. Related items can be shown on such catalogue to attract customers and promote the business.

4. Association also plays important role in **medical treatments** for e.g. association between various diseases which are likely to be occurring simultaneously or in a sequence [18] can help in treatments patients.

These association rules can be applied to business functions and sales can be promoted let's understand the concept with following two examples

In the first example owner of a retail store wants to grow up his store sales by offering various offers to his customers. For the same purpose he wants to conduct a cart\basket data survey. Here the basket data analysis helps him finding out the associated items which are frequently bought together by a customer like e.g. bread and butter are bought together frequently in a single purchase. Now the store owner can make various decisions depending on this basket data analysis like "What will be the sequence of items placement in store" [20]. He can decide what offer to be given on items if purchased together e.g., if item bread is purchased customer will get some discount on butter. This type of analysis on items bought together will definitely increase the sales.

In second example a bank marketing manager wants to conduct an analysis that which product or services of bank are bought together or sequentially. The purchase data of bank customers are presented as a transaction database so that the sequentially bought products and services can be identified easily and applied for sales promotion [20, 21]. After a proper transaction database analysis and based on rule generated if personal loan and credit card are frequently bought then bank can create an offer for his new customers that every savings customer can be cross sold with a personal loan and credit card. This will definitely drive bank sales and revenue.

CONCLUSION

Frequent pattern mining is a strong approach to grow the business with application of various association rules extracted. It is easy to understand and implement. Apriori algorithm is a good method to extract the association rules on some transactional database, but it has limitations like if the database has large number of items, its computation rate will be very high.

References

[1] R. Srikmlt and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In Proc. 5th Int. Conf. Extending Database Technology (EDBT), pages 3 17, Avignon, France, March 1996.

[2] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In Proc. I996 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'96), pages 1 12, Montreal, Canada, June 1996.

[3] H. Mannila, H Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery, 1:259 289, 1997.

[4] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In Proe. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'O0), Dallas, TX, May 2000.

[5] C. C. Aggarwal, P. S. Yu. Online Generation of Association Rules, *ICDE Conference*, 1998.

[6] C. C. Aggarwal, P. S. Yu. A New Framework for Itemset Generation, *ACM PODS Conference*, 1998.
[7] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining Frequent Patterns with Counting Inference. *ACM SIGKDD Explorations Newsletter*, 2(2), pp. 66–75, 2000.
[8] R. J. Bayardo Jr. Efficiently mining long patterns from databases, *ACM SIGMOD Conference*, 1998.
[9] C. Borgelt, R. Kruse. Induction of Association Rules: Apriori Implementation, *Conference on Computational Statistics*, 2002. URL
[10] J.-F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets: A Condensed Representation of Boolean data for the Approximation of Frequency Queries. *Data Mining and Knowledge Discovery*, 7(1), pp. 5–22, 2003.
[11] H. Peng, "Discovery of Interesting Association Rules Based on Web Usage Mining," International Conference, 2010.
[12] R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules," In Proc. VLDB, pp. 487-499, 1994.
[13] Maragatham, M. Lakshmi, and G. Maragatham, "A Recent Review on Association Rule Mining," in Indian Journal of Computer Science. vol. 2, 2012.
[14] X. Wang, Y. Ouyang, X. Hu, and Y. Zhang, "Discovery of User Frequent Access Patterns on Web Usage Mining," in Computer Supported Cooperative Work in Design Proceedings 8th IEEE International Conference. vol. I, pp. 765-769, 2004.
[15] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," in In Proc. Of the ACM STGMOD Int. Conf. On management of Data, Dallas, Texas, USA, 2000.
[16] H.Yao, H.J.Hamilton, C.J.Butz, A foundation approach to mining item set utilities from databases,in:Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida , 2004,pp.482-486
[17] H.Yao,H.J.Hamilton ,Mining item set utilities from transaction databases, in Data and Knowledge Engineering 59(2006) pp.603-626
[18] Cheng-Wei Wu, Souleymane Zida1, Vincent S. Tseng, FHM: Faster High-Utility Item set Mining using Estimated Utility Co-occurrence Pruning ,21st International Symposium for intelligent systems, (ISMIS 2014), Springer LNAI pp 83-92.
[19] Wei Wu, Bai-EnShie, and Philip S. Yu ,UP-Growth: An Efficient Algorithm for High Utility Item set Mining ,KDD'10, July 25–28, 2010, Washington, DC, USA.
[20] Jieh-shanYeh, Yu-chiang Li , Chin-chen Chang, Two-phase algorithms for a novel utility-frequent model,in : Emerging Trends in Knowledge Discovery and Data Mining, Lecture notes in Computer Science, Vol. 4819/2007, pp. 433-444.
[21] R. Uday Kiran, P.Krishna Reddy, Discovering diverse-frequent patterns in transactional databases, International Conference COMAD-2011, December 19-21 2011.