International Scientific Journal of Engineering and Management
Volume: 03 Issue: 05 | May – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

Deciphering Patterns in a small scale Case Analysis Study of the Apriori Algorithm in Market Basket Analysis using machine learning tools

Dr R.Naveenkumar, Associate Professor, Department of Computer Science and Engineering, Brainware University, Kolkata, West Bengal, E-Mail: rnaveenkumarooty@gmail.com

Md Aisaifi, Department of Computer Science and Engineering, Brainware University, Kolkata, West Bengal.

Abstract:

Analysis of Transaction Database Using Apriori Algorithm. This study examines the use of the Apriori algorithm for analysing transaction databases. The Apriori algorithm is a fundamental technique in data mining that allows for the efficient discovery of frequent patterns and association rules in large datasets. The Apriori algorithm employs a two-step approach. Initially, it identifies frequent items in the database based on a user-defined minimum support threshold. Subsequently, it generates association rules that describe relationships between these frequent items based on metrics such as confidence and lift. This paper provides an in-depth explanation of the Apriori algorithm, emphasizing its strengths and limitations. Additionally, it presents various applications of the Apriori algorithm in real-world scenarios, such as shopping cart analysis, cross-selling and upselling, and customer segmentation. The significance of this study lies in its comprehensive analysis of the Apriori algorithm and its practical relevance in diverse data mining tasks. It serves as a valuable resource for researchers, practitioners, and anyone seeking to understand the fundamentals and applications of association rule mining.

Keywords: Datamining, OLTP, Apriori algorithm, Item sets and Market Basket Analysis

Introduction:

A transactional database is a transactional data store or OLTP (Online Transaction Processing) database which is a type of database optimized for handling and managing transaction-oriented workloads[1]. In this type of database, transactions are regarded as the single units of work or operations that modify the database's state. These transactions involve actions such as inserting, updating, deleting, or retrieving data from the data warehouse. OLAP, or Online Analytical Processing, refers to a technology used for querying and analysing data to support business intelligence activities [2]. While OLAP is not a physical entity or service specific to Kolkata, it is a tool widely used in various industries and businesses across the globe, including those in Kolkata [3]. In Kolkata, businesses ranging from retail to finance, healthcare to manufacturing, utilize OLAP

ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

solutions to gain insights from their data, make informed decisions, and drive growth. For example, a retail chain in Kolkata might use OLAP to analyze sales data across different stores, identify trends, and optimize inventory management [4]. Similarly, a financial institution might use OLAP to analyze customer transaction data, detect patterns of fraud, and improve risk management strategies. While OLAP itself is not unique to Kolkata, the application of OLAP technology within businesses in the city contributes to their efficiency, competitiveness, and ability to adapt to changing market dynamics. As Kolkata continues to evolve as a hub for commerce and industry, OLAP remains a valuable tool for businesses seeking to leverage data-driven insights for success [5]. The Apriori algorithm is a basic technique in data mining and association rule learning, primarily used for finding patterns in transactional databases. It is widely used in market basket analysis, where the task is to identify relations between different products bought at once by the customers. It is named after the idea of "prior knowledge" which means “priori" in Latin, the Apriori algorithm displays the concept of efficiency in the generation of association rules to work with transactional data [6]. The Apriori algorithm is used to cope up with the challenge of identifying frequent item sets within a huge dataset[7]. The algorithm works in a two-step approach: first, it identifies all frequent item sets, and then it generates association rules from these item sets based on user-specified criteria such as support, confidence, and lift from the database [8]. The basic idea behind the Apriori algorithm is the "apriori property," which states that if an item is frequent, then all of its subsets must also be frequent [9]. Leveraging this property, the algorithm efficiently prunes the search space, reducing computational complexity [10].

Approach to identifying unique item sets:

Generate Candidate Item sets: The algorithm starts by identifying all unique items present in the dataset and creates initial candidate item sets consisting of individual items (1-item sets). It then iteratively generates larger candidate item sets by joining frequent item sets from the previous iteration. Calculate Support: Once the candidate item sets are generated, the algorithm scans the transaction database to count the occurrences of each candidate itemset. The support of an itemset is defined as the proportion of transactions in which the item set appears.Prune Infrequent Item sets: Candidate item sets with support below a user-defined minimum support threshold are pruned from further consideration. This pruning step exploits the apriori property, ensuring that only potentially frequent item sets are retained for subsequent iterations. Generate Association Rules: After identifying all frequent item sets, the algorithm generates association rules based on user-defined metrics such as confidence and lift. Association rules express relationships between items, indicating the likelihood of one item being purchased given the presence of another item.

Primarily used for association rule mining:

The Apriori algorithm is primarily used for association rule mining in transactional databases.

ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

Market Basket Analysis: One of the most common applications of the Apriori algorithm is in market basket analysis. Supermarkets and retailers use it to understand the relationships between different products that customers tend to buy together. By identifying frequent item sets and association rules, retailers can optimize product placement, promotions, and recommendations to increase sales and customer satisfaction. Cross-Selling and Upselling: E-commerce platforms leverage the insights gained from association rule mining to implement cross-selling and upselling strategies. By recommending related or complementary products based on past purchase patterns, companies can increase the average order value and enhance the shopping experience for customers. Customer Segmentation: Apriori algorithm can also be used for customer segmentation by identifying groups of customers with similar purchasing behaviours. This information enables businesses to tailor marketing campaigns, promotions, and product offerings to specific customer segments, leading to more targeted and effective marketing strategies. The Apriori algorithm is a data mining technique used for finding frequent item sets and association rules from transactional data. It's commonly used in market basket analysis, which is highly applicable to a small grocery shop in Kolkata. Let's break down the process: Frequent Itemset Generation: The algorithm starts by identifying all individual items and their frequencies in the transaction dataset. Then, it gradually generates larger item sets by combining the frequent item sets found in the previous iteration. The support count of each item set is calculated, representing how often it appears in the transactions.

1. Formula for support count (support) of an item set:
   support(X)=Transactions containing XTotal transactionsSupport(X)=Total transactionsTransactions c ontaining X
2. Here, XX represents an item set.
3. Association Rule Generation: Once the frequent item sets are identified, association rules are generated based on certain criteria, typically support and confidence. Support measures how frequently the items in the rule co-occur, while confidence measures the reliability of the rule.
4. Formula for confidence of an association rule:
   Confidence($X \rightarrow Y$)=Support($X \cup Y$)Support(X)Confidence(X→Y)=Support(X)Support(X∪Y)
5. Here, XX and YY are item sets.
6. For a small grocery shop in Kolkata, the process of applying the Apriori algorithm involves the following steps:Data Collection: Gather transactional data, which includes records of items purchased together in each transaction. For example, a transaction may consist of items like rice, lentils, and vegetables. Frequent Itemset Generation: Apply the Apriori algorithm to identify frequent item sets

ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

from the transactional data. Determine the minimum support threshold, which defines the minimum frequency required for an item set to be considered frequent. Association Rule Generation: Once frequent item sets are identified, generate association rules based on support and confidence thresholds. These rules reveal associations between items that frequently occur together in transactions. Analysis and Action: Analyze the generated association rules to gain insights into customer purchasing behavior. Use these insights to optimize product placement, inventory management, marketing strategies, and promotions in the grocery shop.By implementing the Apriori algorithm and analyzing the results, a small grocery shop in Kolkata can make data-driven decisions to enhance its operations and improve customer satisfaction

Transaction Id	Items
1	{milk, bread, butter}
2	{bread, butter, eggs}
3	{milk, bread, eggs}
4	{milk, butter}
5	{bread, eggs}
6	{bread, butter}
7	{milk, bread, butter, eggs}
8	{milk, bread}
9	{bread, butter, eggs}
10	{bread, butter, eggs}

Table 1: Data item set

Identify Unique Items:

Unique items: {bread, milk, eggs, butter, cheese}

Generate Candidate Item sets:

1-item sets: {bread}, {milk}, {eggs}, {butter}, {cheese}

Calculate Support:

Itemset	Count(support)
Bread	8
Milk	5

ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

Eggs	2
Butter	5
Cheese	5

Table 2: Generate Candidate Item sets of Higher Order:2

Item Sets	Count(support)
{bread, milk}	4
{bread, butter}	5
{bread, cheese}	4
{milk, butter}	2
{milk, cheese}	2
{butter, cheese}	4
{bread, eggs}	2

Table 3: Generate Candidate Item sets of Higher Order:3

Item sets	Count(support)
{bread, milk, butter}	3
{bread, milk, cheese}	3
{bread, butter, cheese}	4
{milk, butter, cheese}	2

Table 4: Count (Support) frequent set 2

Frequent 2-item sets:

{bread, milk}, {bread, butter}, {bread, cheese}, {butter, cheese}, {bread, milk, butter}, {bread, milk, cheese}, {bread, butter, cheese}

Frequent 3-item sets:

{bread, milk, butter}, {bread, milk, cheese}, {bread, butter, cheese}

Generate Association Rules:

For each frequent itemset, generate association rules and calculate confidence.

Example rules:

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 05 | May – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

{bread, milk} => {butter}: confidence = support ({bread, milk, butter}) / support ({bread, milk}) = 3/4 = 0.75

{bread, butter} => {cheese}: confidence = support ({bread, butter, cheese}) / support ({bread, butter}) = 4/5 = 0.8

This is for the daily sales of a shop in Kolkata for a month

Itemset	Count(support)	Price	Quantity	Amount
Bread	8	20	50	1000
Milk	5	31	20	620
Eggs	2	5	45	225
Butter	5	80	84	6720
Cheese	5	100	96	9600

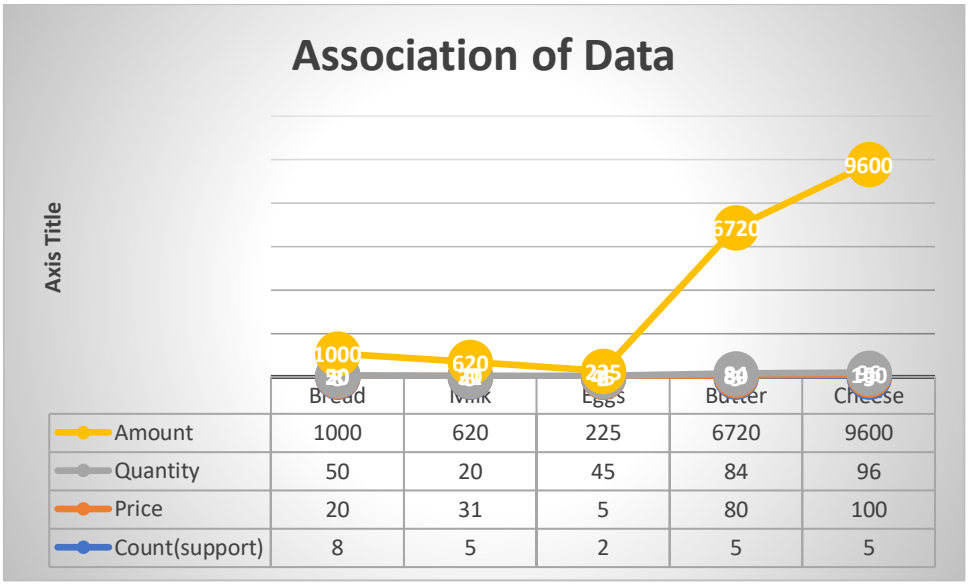


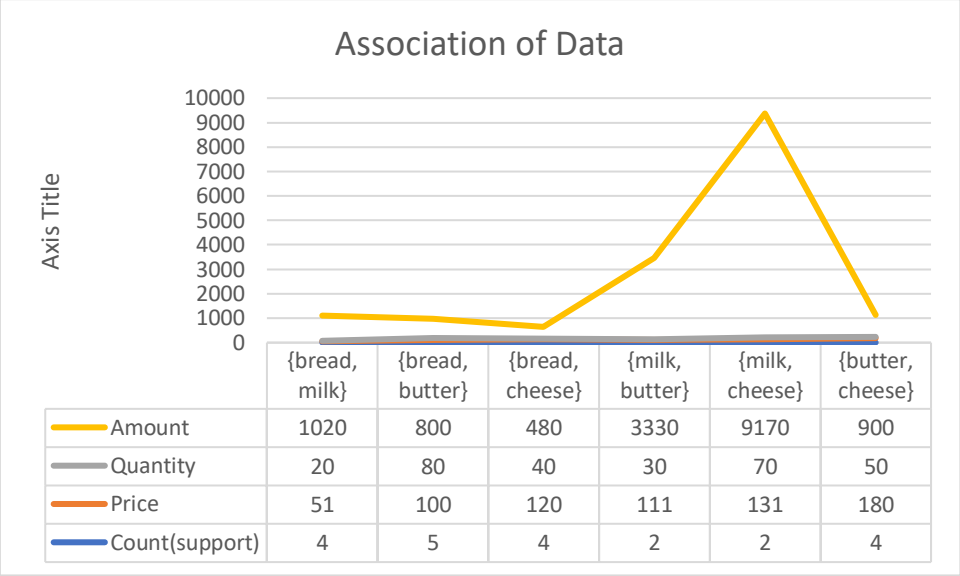
Figure : 1

The association of data of two itemset

Item Sets	Count(support)	Price	Quantity	Amount
{bread, milk}	4	51	20	1020
{bread, butter}	5	100	80	800
{bread, cheese}	4	120	40	480
{milk, butter}	2	111	30	3330
{milk, cheese}	2	131	70	9170
{butter, cheese}	4	180	50	900

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 05 | May – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

The association of data in three data set



	{bread, milk}	{bread, butter}	{bread, cheese}	{milk, butter}	{milk, cheese}	{butter, cheese}
Amount	1020	800	480	3330	9170	900
Quantity	20	80	40	30	70	50
Price	51	100	120	111	131	180
Count(support)	4	5	4	2	2	4

Figure : 2

Item Sets: Each row represents a combination of items sold together in transactions.

Count (Support): This column indicates the frequency of occurrence of each itemset in the dataset. For example, the item {bread, milk} appeared in 4 transactions, {bread, butter} in 5 transactions, and so on.

Price: This column displays the price associated with each itemset combination. It represents the total price of all items in the respective itemset. For instance, the price of {bread, milk} is 51, {bread, butter} is 100, and so forth.

Quantity: This column denotes the quantity of each itemset combination sold in transactions. It reflects the total number of times all items in the itemset were sold together. For example, for the itemset {bread, milk}, a total quantity of 20 units was sold across 4 transactions.

Amount: This column indicates the total revenue generated from each itemset combination. It is calculated by multiplying the price of the itemset by the quantity sold.For instance, the amount generated from selling {bread, milk} is 1020 (51 * 20), {bread, butter} is 800 (100 * 80), and so on.

Elaborating on this dataset provides insights into the sales patterns and revenue generated from different combinations of items sold together. Analyzing this information can help the shopkeeper understand

ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

customer preferences, optimize pricing strategies, and identify opportunities for promoting bundled products or cross-selling items. Additionally, it facilitates decision-making related to inventory management and stocking levels based on the popularity and profitability of various item combinations

Item sets	Count(support)	Price	Quantity	Amount
{bread, milk, butter}	3	131	6	786
{bread, milk, cheese}	3	151	8	1208
{bread, butter, cheese}	4	200	20	4000
{milk, butter, cheese}	2	211	30	6330

According to the above data set, If the shopkeeper of the small shop in Kolkata sells where he has more number of item sets out of we have taken 3 most frequently sold item set for calculation.

{Cheese}

{milk, cheese}

{milk, butter, cheese}

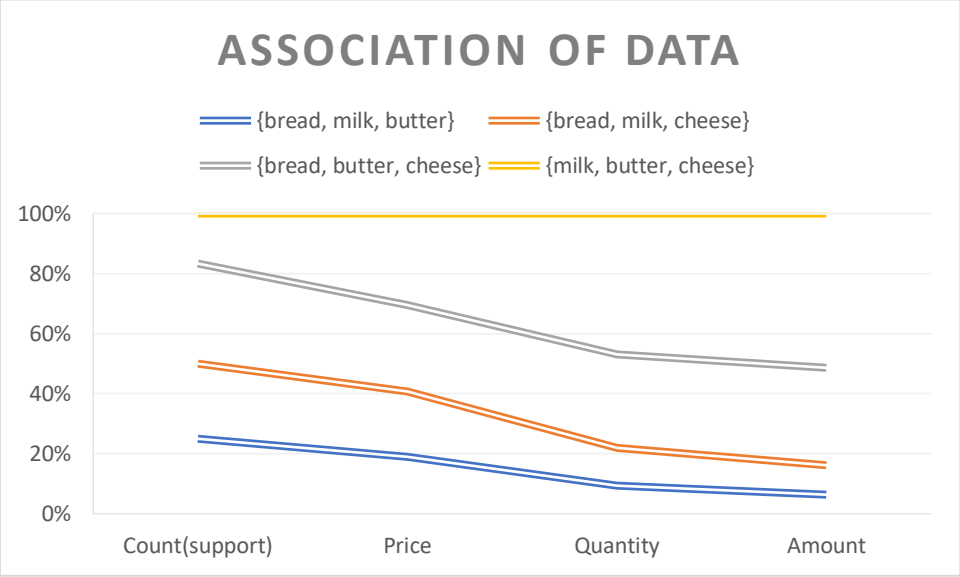


Figure : 3

It will be the most profitable way to make money with lesser investment of time and energy .

ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

the scenario based on the given dataset where the shopkeeper of a small shop in Kolkata wants to focus on selling the most frequently sold item sets out of the three identified:

Cheese: This item consists of only one item, cheese. The shopkeeper can analyze the sales frequency and patterns associated with cheese alone. Elaboration: The shopkeeper can track the sales volume of cheese over time, identify peak periods of demand, and adjust stocking levels accordingly. Additionally, they can explore potential correlations between cheese sales and other factors such as seasonality or promotional activities.

Milk, Cheese: This item includes two items, milk and cheese, sold together in transactions. The shopkeeper can examine the association between milk and cheese sales and how their combined purchase impacts overall sales. Elaboration: By analyzing the co-occurrence of milk and cheese in transactions, the shopkeeper can identify complementary products and potentially bundle them together for promotions or special offers. Understanding the relationship between these items can also inform pricing strategies and inventory management decisions.

Milk, Butter, Cheese: This item comprises three items, milk, butter, and cheese, commonly sold together. The shopkeeper can explore the dynamics of selling these three items as a bundle and assess their collective impact on sales. Elaboration: Selling milk, butter, and cheese together as a bundle can create value for customers seeking convenience or looking to purchase all three items in one transaction. The shopkeeper can analyze the profitability of offering this bundle compared to selling the items individually and adjust the pricing or promotion strategies accordingly.

Conclusion:

In conclusion, the Apriori algorithm stands as a powerful tool for discovering association rules within large datasets, particularly in the domain of market basket analysis. Through its iterative approach of generating candidate item sets and pruning infrequent ones, Apriori efficiently identifies significant associations between items. Its ability to handle large datasets and scalability makes it widely applicable in various industries, ranging from retail to healthcare and beyond. However, while Apriori provides valuable insights into item co-occurrence patterns for bread, milk, and butter, it does have its limitations. Its performance may degrade with very large datasets or high-dimensional data, and it may produce numerous redundant or irrelevant rules without additional refinement. Additionally, the algorithm assumes independence between items within transactions, which may not always hold true in real-world scenarios. To maximize the utility of

ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

Apriori-generated association rules, it's essential to complement the algorithm with post-processing techniques such as rule filtering, pruning, and evaluation metrics like support, confidence, and lift. This ensures that only the most meaningful and actionable rules are retained for decision-making purposes. Despite its challenges, the Apriori algorithm remains a foundational technique in data mining and association rule learning, offering valuable insights into transactional data and facilitating informed decision-making in various domains. As datasets continue to grow in size and complexity, ongoing research and innovation in association rule mining techniques will further enhance the capabilities and applicability of algorithms like Apriori . Overall, by focusing on the most frequently sold item sets like {Cheese}, {Milk, Cheese}, and {Milk, Butter, Cheese}, the shopkeeper can gain insights into customer preferences, optimize product offerings, and enhance the shopping experience for customers in Kolkata. Analyzing these item sets can inform strategic decisions related to inventory management, pricing, and marketing efforts, ultimately driving sales and profitability for the small shop

Reference:

[1] Achmad Nur Cahyon, "Implementation of Data Mining of Apriori Algorithms on Motorcycle Spare Parts Sales in Ahas Putra Motor", 2016

[2] Al-maolegi, Mohammed, and Bassam Arkok. A N I MPROVED A PRIORI A LGORITHM The International Conference on Computer Science and Applied Mathematic IOP Conf. Series: Journal of Physics: Conf. Series 1255 (2019) 012057 IOP Publishing doi:10.1088/1742-6596/1255/1/012057 8 FOR. Vol. 3, no. 1, 2014, pp. 21–29.

[3] Bhandari, Akshita, et al. “Improvised Apriori Algorithm Using Frequent Pattern Tree for Real Time Applications in Data Mining.” Procedia - Procedia Computer Science, vol. 46, no. Icict 2014, Elsevier Masson SAS, 2015, pp. 644–51, doi:10.1016/j.procs.2015.02.115.

[4] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. P. Windarto, “Application of Classification Method C4 . 5 on Selection of Exemplary Teachers,” in IOP Conference Series, 2018, pp. 1–6.

[5] Desti Fitriati, "Implementation of Data Mining to Determine the Combination of Goods Promotion Media Based on Customer Purchase Behavior Using Apriorí Algorithms", Annual Research Seminar, 2016, 2 (1)

[6] Domma Lingga, "Application of Apriori Algorithms in Predicting Book Inventory at the Dwi Tunggal Tanjung Morawa High School Library", Information and scientific technology, 2016, XI (1)

[7] Dutt, Shalini, et al. “An Improved Apriori Algorithm Based on Matrix Data Structure.” Global Journal of Computer Science and Technology: C Software & Data Engineering, vol. 14, no. 5, 2014, pp. 1–5.

ISSN: 2583-6129
DOI: 10.55041/ISJEM01603

[8] Fadlina "Data Mining For Analysis of Street Crime Levels with Apriori Association Rule Algorithm Apriori Method", Scientific Information and Technology (CORE), 2014, III (1)

[9] S. Sudirman, A. P. Windarto, and A. Wanto, "Data Mining Tools | RapidMiner : K-Means Method on Clustering of Rice Crops by Province as Efforts to Stabilize Food Crops In Indonesia," IOP Conference Series: Materials Science and Engineering, vol. 420, no. 12089, pp. 1–8, 2018.

[10] Liu, Xiyu, et al. "An Improved Apriori Algorithm Based on an Evolution-Communication Tissue-Like P System with Promoters and Inhibitors." Hindawi Discrete Dynamics in Nature and Society, vol. 2017, 2017, pp. 1–12

