JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 798 - 803
https://publishoa.com
ISSN: 1309-3452

Association Rule Mining Using Retail Market Basket Dataset by Apriori and FPGrowth Algorithms

Dr. Harvendra Kumar Patel, Prof. (Dr.) K. P. Yadav
harvendra.patel82@gmail.com, drkpyadav732@gmail.com
Mats University, Chhattisgarh, India

ABSTRACT
Data mining is a method for dealing with large amounts of data. It takes data from a large dataset and extracts meaningful information. A suite of algorithms has been developed to extract meaningful information from big datasets. Apriori, ECLAT, FPGrowth, and others are examples of such algorithms. These algorithms are mostly used to identify the frequent itemsets. There are two models and eight functions in data mining, and each model has four different functions. In this study, we will employ one Apriori technique and the other, the FPGrowth algorithm, to find frequent itemsets. These methods operate on the same dataset in different ways, extracting the same frequent itemsets but with varied execution times. The remainder of this work is organized logically. The rest of the work is arranged in the following manner: The first segment begins with an introduction. The second segment is a review of related literature. The third segment looks into the foundational ideas. The outcome and analysis are depicted in Section 4. Finally, Section 5 brings this paper to a close by discussing the implications of our work for future research endeavors.

Keywords: Data Mining, Association Rule Mining, Association Rule Mining Measures, Apriori Algorithm, FPGrowth Algorithm.

1. Introduction

Data is being collected from a variety of sources, including mobile devices, IoT devices, sensor devices, the internet, social media sites, audio, video, and so on. It is quite challenging to handle large datasets. Today, data size is the most important challenge. Data mining is a good tool for managing big amounts of data. Data mining is a technique for retrieving valuable data from big datasets. As demonstrated in fig. 1, data mining incorporates a range of subfields, notably machine learning, statistics, database systems, and so forth. Agarwal, Lamilinsky, & Swamy (1993), Agarwal & Srikanth (1994) proved that huge databases, including price-based information, buyer information, deal records, and so on, include many associations. For massive data, they devised the Apriori approach (Agrawal and Srikant, 1994). The Apriori algorithm, ECLAT algorithm, FPGrowth algorithm, and many others are used to operate on the obtained dataset in data mining [1]. The machine learning paradigm covers the Apriori and FPGrowth algorithms, which use unsupervised learning methods. Machine learning techniques that are unsupervised identify hidden patterns in datasets. It finds connections and links between datasets. Data mining entails a variety of domains, as indicated in Figure 1.

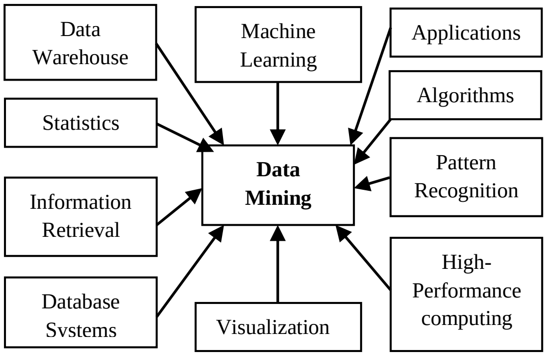

Fig 1: Different subfields of Data Mining

2. Literature Review

Agarwal et al. [3] proposed the mechanism of ARM in 1993. We have taken the most popular problem, Market Basket Analysis (MBA), and utilized it to figure out the association rule in this paper. This analysis's conclusion is based on the customer's purchasing mechanism. What are the chances of a consumer buying milk if they buy bread and butter? This mechanism informs retailers about a customer's proclivity to purchase a product from the market. Bagui and Dhar [4] suggest a mechanism for mining positive and negative association rules in the Hadoop MapReduce environment. The researchers developed a classic ARM method that simply looks for positive association rules. Positive rule mining has applications in a variety of domains, including stock market analysis, weblog mining, medical diagnostics, customer market analysis, and bioinformatics [5]. The negative ARM, on the other hand, has the advantage that if two itemsets are negatively connected, one item will rise and the other will fall. In the field of ARM, it has a wide range of applications, such as crime data analysis and healthcare analysis.

For frequent pattern mining of market basket datasets, Jeff Heaton [6] proposed a naive method. The naive method first creates all potential itemsets, then counts their support. After counting the support, it discards those itemsets which have low threshold values. On the Tang dataset, which is a researcher platform, Yi Zeng et al. [7] used three techniques. The algorithms Painting-Growth, N-Painting Growth, and FPGrowth have been applied to the Tang dataset. Over the Tang dataset, each method iterated 20 times. Following the application of the algorithms to the Tang dataset, the results are compared in terms of execution time. According to Jiao Yabing [8], the Boolean algorithm originates from the Apriori algorithm for the frequent mining of itemsets. The most important thing behind this is that the subsets of frequent itemsets will be frequent itemsets, and if the superset is non-frequent, then itemsets will be non-frequent. Nasreen et al. [9] have discussed pattern recognition for data mining. Several algorithms have been analyzed in this research work for finding frequent patterns. A comparative study has been conducted over the various algorithms named as the Priori algorithm, Frequent Pattern (FP) Growth algorithm, Rapid Association Rule Mining, ECLAT algorithm, and Associated Sensor Pattern Mining of Data Strean (ASPMS) over a large dataset for finding the patterns. These algorithms can be used for various real-world problems like market basket analysis, and promotion of various types of products in the market. Christian Borgelt [10] proposed the implementation of three data mining algorithms: Apriori, Eclat, and Relim. The FPGrowth algorithm represents a prefix tree of transactions for the provided dataset. It is considered a recursive elimination scheme. In recursive elimination, first of all, eliminate all those transactions that are non-repetitive, or non-frequent itemset in the dataset. After that, choose only those transactions that have the least frequent items in the dataset. Liu et al., [11] proposed a fast Apriori known as ECTPPI-Apriori. It is used for large datasets because the main motive is to find the frequent itemsets. The researchers worked on the traditional market basket dataset.

3. Preliminaries

Association rule mining is considered as an if-then X→Y relationship. In this, if itemset X is being purchased by the customer, then the probability of itemset Y being selected in the same transaction is ascertained. The main purpose of this algorithm is to find the relationship between the objects in the transactional dataset [13,14,15,16].

a) **An item** is a field in a transactional database. The item means it is one. In the binary system, if the item is present, it is represented by 1 and if the item is not present, it is represented by 0.

b) **NULL transaction:** A transaction that does not have any items is known as a null transaction. If a dataset has null transactions, then association rules cannot be found.

c) **Itemset: It** is a collection of items. The itemset of n items is represented as $\{I_1, I_2, I_3,.., In\}$.

d) **Frequent itemset:** If the support for the set of elements is greater than or equal to the minimum support threshold (min_supp), then the set of elements is said to be a frequent itemset.

e) **Support (s): - The support** is defined as the number of transactions contained in a total transaction. It is also known as an itemset's count or frequency in a dataset [12].

$$Support\ (X \rightarrow Y) = P(XY)$$
$$n(X \cup Y)/N$$

f) **Confidence (c)**: - Confidence is the percentage of transactions with itemset X in database D that also contain itemset Y. Confidence is a conditional probability. The mathematical formula for confidence is given as:

$$Confidence\ (X \rightarrow Y) = P(Y/X) = s(X \cup Y)/s\ (X)$$
$$= fre(X,Y)/fre(X) = n(X \cup Y)/n(X)$$

The rule with the highest support is given preference over the rules of equal confidence. The rationale is that confidence estimates tend to be more reliable. The range of confidence is 0 to 1.

g) **Lift/Interest (I): -** It is the ratio of the confidence in the rule and the expected confidence in the rule. Lift/Interest is used to measure frequency X and Y together if both are statistically independent of each other [8, 9]. The lift of rule X → Y is defined as,

$$lift(X \rightarrow Y) = lift(Y \rightarrow X) = P(X\ and\ Y)/(P(X)P(Y)) = conf(X \rightarrow Y)/sup(Y) = conf(Y \rightarrow X)/sup(X)$$

$$lift(X \rightarrow Y) = Confidence\ /Expected\ confidence = Confidence(X \rightarrow Y)/\ Support(Y)$$
$$= Support(X \rightarrow Y)/(\ Support(X) * Support(Y))$$
$$= fre(X,Y) * N/(fre(X) * fre(Y))$$

h) **Conviction (Conv): -** A high conviction value means that the outcome is highly dependent on the predecessor. For example, in the case of a complete confidence result, the denominator becomes 0 (due to 1 - 1) where the sentencing score is defined as "inf". Similar to lift, if things are independent, the conviction is 1. Legal sentences are defined as:

$$conv(X \rightarrow Y) = (1 - supp\ (Y))/(1 - conf(X \rightarrow Y))$$
$$conv(Y \rightarrow X) = (1 - supp\ (X))/(1 - conf(Y \rightarrow X))$$

So, conviction is not symmetric.

Table 1: Range of measures

Name	Equation	Feasible values
Support	P_{xy}	[0,1]
Confidence	$\frac{P_{xy}}{P_x}$	[0,1]
Lift	$\frac{P_{xy}}{P_x * P_y}$	[0,1]
Conviction	$\frac{P_x * P_{\bar{y}}}{P_{x\bar{y}}}$	$\left[\frac{1}{n}, \frac{n}{4}\right]$

4. Methodology and Result

Association rule mining is a two-step process. The first step is to discover frequent itemsets and the second step is to generate the association rule. Therefore, the Apriori algorithm and the FP growth algorithm are used to take out frequent itemsets, and these frequent itemsets are used to identify patterns in the dataset. The grocery dataset is used in this paper. Fig 2 represents the working methodology of the method. Apriori and FPGrowth algorithms are applied to datasets one by one to fetch the insides. This research used the grocery dataset, which has 38765 transactions and three attributes, namely member_number, date, and item description. The dataset has 167 unique items. A total of 38765 items were sold in 728 days throughout 12 months, an average of 53.25 items sold daily.

Apriori Algorithm: The Apriori algorithm is applied to the grocery dataset. This is a two-step approach. The first phase is join-step and the second phase is prone. Join the itemset to the next ordered itemset in the join step and remove non-frequent itemsets in the prone step. Frequent itemset generation is a part of an algorithm (Apriori, Eclat, FPGrowth, etc.). Level by level traversing the itemset is done by the Apriori algorithm. This algorithm uses the BFS technique. In the first iteration, it generates a 1-frequent itemset, in the second iteration, it generates a 2-frequent itemset, in the third iteration, it generates a 3-frequent itemset, and so on. The maximum number of iterations required by this algorithm is I_{max}+1, where I_{max} is the largest size of the frequent itemsets.

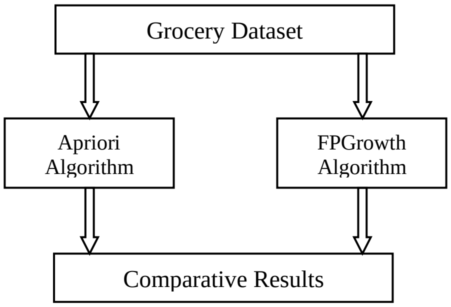

Fig 2: Comparative results

FP-Growth: The FPGrowth algorithm is used to mine frequent itemsets from a dataset. This is an improved version of the Apriori method. Here, frequent itemsets are generated without coordinating candidate generation. The entire database is displayed as a tree, so it is known as a "frequent pattern tree." The whole tree is the association that provides coordination between objects. The entire database is segmented through frequent items, and these segmented items are known as pattern segments. These patterns are analyzed to reduce the search for itemsets. The following steps should be considered for the execution of the FPGrowth algorithm:

Step 1: The entire database is scanned to detect occurrences of an itemset.
Step 2: The FP tree is designed, and the root node is considered null.
Step 3: Scan the database and inspect the transactions.
Step 4: The transaction is checked in the database once again to find out whether the itemsets are in descending order or not. Itemsets should be in a tree-like connected form.
The grocery dataset has 167 unique items and three attributes (features), namely Member_Number, Date, and Item_Description, in a transactional dataset. The transactional dataset has 38765 items that were sold in 728 days throughout 12 months, i.e., an average of 53.25 items sold daily.

1	Member_Number	38765 non-null
2	Date	38765 non-null
3	Item_Description	38765 non-null

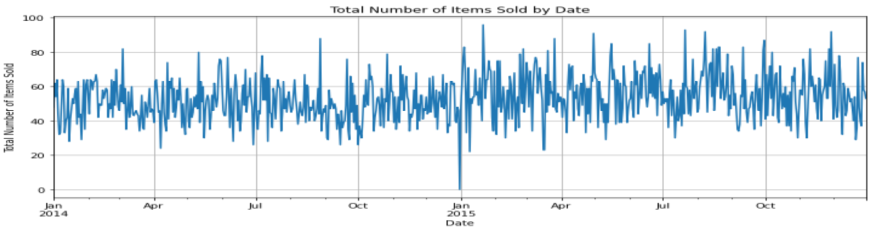

Fig 3: Total number of items sold by date

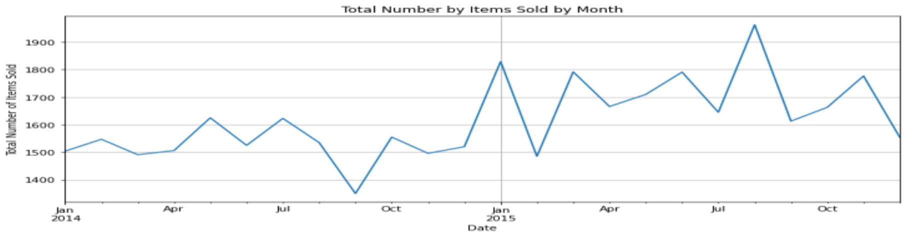

Fig 4: Total number of items sold by the month

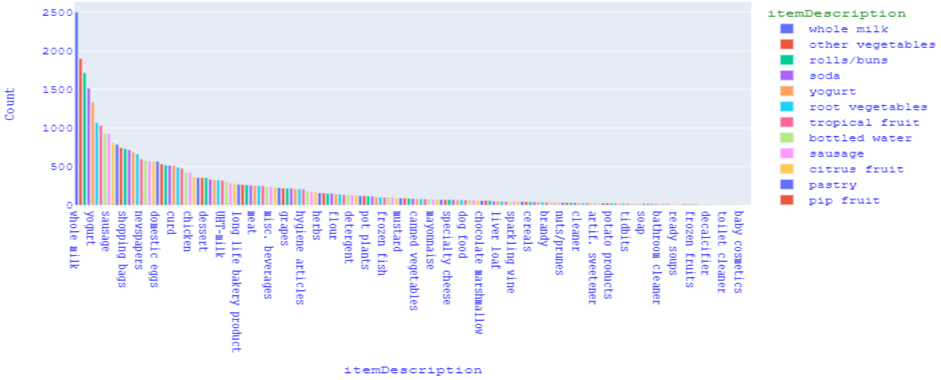

Fig 5: Items description

Association rules by Apriori Algorithm

Index	antecedents	Consequents
0	frozenset({'UHT-milk'})	frozenset({'other vegetables'})
1	frozenset({'UHT-milk'})	frozenset({'whole milk'})
2	frozenset({'beef'})	frozenset({'whole milk'})
3	frozenset({'berries'})	frozenset({'other vegetables'})
4	frozenset({'berries'})	frozenset({'whole milk'})
5	frozenset({'beverages'})	frozenset({'other vegetables'})
6	frozenset({'beverages'})	frozenset({'soda'})
7	frozenset({'beverages'})	frozenset({'whole milk'})
8	frozenset({'bottled beer'})	frozenset({'other vegetables'})
9	frozenset({'bottled beer'})	frozenset({'whole milk'})

Association rules by FPGrowth Algorithm

index	Antecedents	Consequents
0	frozenset({'yogurt'})	frozenset({'whole milk'})
1	frozenset({'yogurt', 'whole milk'})	frozenset({'other vegetables'})
2	frozenset({'yogurt', 'other vegetables'})	frozenset({'whole milk'})
3	frozenset({'rolls/buns', 'yogurt'})	frozenset({'whole milk'})
4	frozenset({'yogurt', 'whole milk'})	frozenset({'rolls/buns'})
5	frozenset({'sausage'})	frozenset({'whole milk'})
6	frozenset({'yogurt', 'whole milk'})	frozenset({'sausage'})
7	frozenset({'yogurt', 'sausage'})	frozenset({'whole milk'})
8	frozenset({'whole milk', 'sausage'})	frozenset({'yogurt'})
9	frozenset({'rolls/buns', 'sausage'})	frozenset({'whole milk'})

The execution times of both algorithms are 19.988207817077637 (Apriori Algorithm) and 15.845721960067749 (FPGwroth Algorithm) respectively. The Apriori algorithm takes more time than FPGrowth for a given grocery dataset.

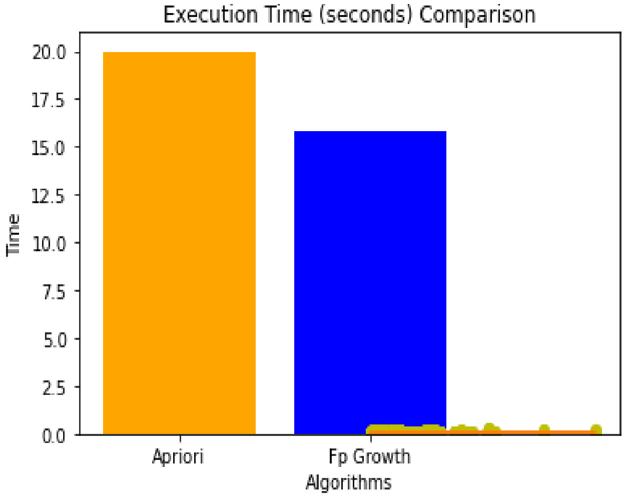

Fig 6: Execution time for both Apriori algorithm and FP Growth algorithm

5. Conclusion

The Apriori algorithm and the FPGrowth algorithm are applied to the grocery dataset. I discovered that the FPGrowth algorithm takes less time than the Apriori approach. Both methods yield the same frequently occurring itemsets and rules. This is because the FPGrowth method only searches the dataset twice, whereas the Apriori approach examines the dataset once for each candidate generation. There is no candidate generation mechanism in the FPGrowth algorithm, whereas the Apriori algorithm uses a candidate generation approach.

References:

1. Hong-Jun Jang et al., "FP-Growth Algorithm for Discovering Region-Based Association Rule in the IoT Environment". Electronics, pp. 1-16, 2021.
2. Changxin Song, "Research of association rule algorithm based on data mining" in the Proceedings IEEE International Conference on Big Data Analysis (ICBDA), IEEE International Conference on Big Data Analysis (ICBDA), pp. 1–4, 2016.
3. Rakesh Agrawal, and Ramakrishnan Srikant "Fast Algorithms for Mining Association Rules in Large Databases" Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, pp. 487-499, 1994.
4. Sikha Bagui and Porbal Chandra Dhar, "Positive and negative association rule mining in Hadoop's MapReduce environment", Journal of Big Data, pp. 2-162019.
5. Stefen Naulaerts, Pieter Meysman, Wout Bittremieux, Trung Nghia Vu, Wim Vanden Berghe, Bart Goethals, and Kris Laukens, "A primer to frequent itemset mining for bioinformatics. Brief Bioinform" 2015, vol. 16, issue. 2, pp. 216-231.
6. Jeff Heaton, "Comparing Dataset Characteristics that Favor the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms" Southeast Con 2016, pp. 1-7, 2016.
7. Yi Zeng, Shiqun Yin, Jiangyue Liu and Miao Zhang, "Research of Improved FP-Growth Algorithm in Association Rules Mining", Hindawi Publishing Corporation Scintific Programming, pp. 1-6, 2015.
8. Jiao Yabing, "Research of an Improved Apriori Algorithm in Data Mining Association Rules", International Journal of Computer and Communication Engineering, vol. 2, issue 1, pp. 25-27, 2016.
9. Shamila Nasreen, Muhammad Awais Azam, Khurram Shehzad and Usman Naeem, "Frequent Pattern Mining Algorithms for Finding Association Frequent Patterns for Data Streams: A Survey", in the proceedings of "5th International Conference on Emerging Uniquitous and Pervasive Networks", pp. 109-116, 2015.
10. Christian Borgelt, "Simple Algorithms for Frequent Item Set Mining", pp. 11-19, 2014.
11. Xiyu Liu, Yuzhen Zhao, and Minghe Sun, "An Improved Apriori Algorithm Based on an Evolution-Communication Tissue-Like P System with Promoters and Inhibitors", Hindawi Discrete Dyanamics in Nature and Society, pp. 1-11, 2017.
12. Ekta Garg and Meenakshi Bansal, "A Survey On Improved Apriori Algorithm", International Journal of Engineering Research and Technology, vol. 2, issue 7, pp. 730-733, 2016.
13. Manpreet Kaura, Shivani Kanga, "Market Basket Analysis: Identify the changing trends of market data using association rule mining", International Conference on Computational Modeling and Security, pp 78-85, 2016.
14. Nilesh Kumar Dokania and Navneet Kaur, "comparative study of various techniques in data mining", International journal of engineering sciences & research technology, pp 202-209, May 2018.
15. Martin Kirchgessner and et al., "Testing Interestingness Measures in Practice: A Large-Scale Analysis of Buying Patterns", IEEE International Conference on Data Science and Advanced Analytics (DSAA), Oct, 2016.
16. Khaled H. Alyoubi, "Association Rule Mining on Customer's Data using Frequent Pattern Algorithm", IJCSNS International Journal of Computer Science and Network Security, VOL.20 No.5, May 2020.