ASSOCIATION RULE MINING ALGORITHMS: A COMPARATIVE REVIEW

Abhishek Sharma[1], Anita Ganpati[2]

[1]*M. Tech Scholar, Department of Computer Science, Himachal Pradesh University, Shimla, India*
[2]*Professor, Department of Computer Science, Himachal Pradesh University, Shimla, India*

---***---

Abstract - *Data mining is defined as process of uncovering beneficial patterns and hidden information from large databases. There are many data mining techniques available for extracting fruitful information and association rule mining is one of them. Association rule mining (ARM) algorithms are quite popular among the researchers all over the world because it makes easier to extract the relation between the items in the form of rules present in the dataset. This paper enlightens and reviews the fundamentals of ARM and compares three pattern mining algorithms i.e. Frequent pattern (FP)-Growth, Equivalence Class Transformation (Eclat) and Apriori algorithm. This study concludes that Eclat and FP-growth preforms better compared to Apriori algorithm in terms of execution time and usage of memory.*

***Key Words*: Data Mining, Association Rule, FP-Growth, Apriori, Frequent Itemsets, Eclat.**

1. INTRODUCTION

The art of finding inconspicuous and beneficial information from large data sources in a proficient way is called Data mining. Data mining is knowledge discovery from data. It is also sometimes referred to as the knowledge mining from data [1]. "Data mining is an integral part of knowledge discovery in databases (KDD), which is the overall process of converting raw data into useful information" [2]. A simple example of data mining is that [3], the data gathered from the supermarket scanners is analysed by the company and this analysis shows that when people are most likely to buy specific products like baby products, bread, milk etc. This ultimately helps the retailers to organize the shelfs of their stores in such way that increases sales and also helps the customers to locate the products easily. Nowadays many well-known companies like google, Netflix, Walmart, amazon, flipkart and many others uses data mining to give suggestions for their products and services to the customers based upon recent history. The five main steps of KDD process are shown in Fig-1 where selection, pre-processing and transformation are the steps where data is prepared for data mining algorithms. In the data mining step different algorithms are applied on the refined data to generate the results. In the last step of evaluation and interpretation interesting trends and patterns are identified and presented to the user using different visualization techniques.

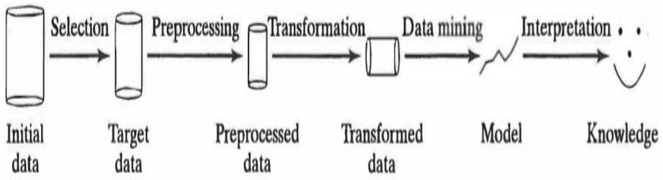


Fig-1: Data mining as a part of KDD process [10]

1.1 Data Mining Tasks and Models

Most commonly used data mining tasks are Classification, Clustering, Regression, Association rule etc. among them Association is the most popular and enthralling area of research in data mining.

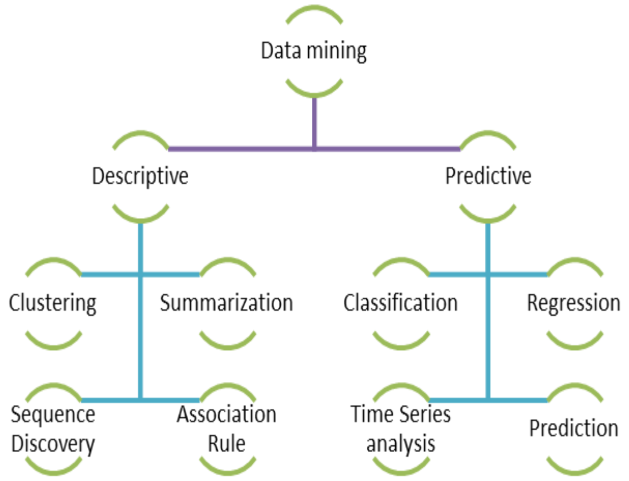


Fig-2: Models and Tasks in Data mining

Predictive Model
A predictive model is use to make predictions about the value of a particular attribute based on the historical data. In this independent variable is used for making the prediction, while dependent variable are the ones whose values will be predicted. Different data mining tasks that comes under the predictive model as shown in Fig-2 are:

- **Classification:** It is one the most important task of data mining and classifies the data into predefined classes. Most popular algorithms are support vector machine and random forest.
- **Regression:** It analyse the relationship that exit between the independent and the dependent variables and also used to forecast the value of a variable when changes are made to other variable.
- **Time series analysis:** In this at different time intervals the value of an attribute will be examined. It is also used to extract statistics and patterns from the data. A popular application of time series analysis is stock market prediction.

- **Prediction:** It is used to detect the unavailable numeric values in the data. Application of prediction includes: flooding, speech recognition etc.

Descriptive Model
The objective of descriptive model is to identify the trends, trajectories and relationships in data. It also provides comprehensive description of the data. Different data mining tasks that comes under the descriptive model are:

- **Clustering:** In clustering grouping is based upon the likeness so that the data present in one cluster is similar to each other while it is dissimilar to the data present in another cluster.
- **Summarization:** It is the generalization and characterization of the data. Summarization is done for finding the high level summarized and compact information about the database.
- **Association rule:** Association rule in data mining represents the relation between items that occur frequently. It helps the retailers to identify the products that are most frequently purchased by the customers.
- **Sequence discovery:** Also known as sequential analysis and used to identify the sequential patterns in data. In this relationships obtained are based on time but patterns are similar to associations in the data. For example, most people who purchase Laptops, may be found to purchase pen drives, speakers etc. within one week [10].

2. OVERVIEW OF ASSOCIATION RULE MINING

Association means relation, connection, union between two or more objects or things. Association rule in data mining represents the relation between items that occur frequently or together. Association rule shows how one thing depends on other like in general store how sales of bread also depends on the sales of milk and should be kept side by side. In association rule mining the association rule are represented in its simplest form as X=>Y where Y is consequent and X is antecedent, example: Milk=>bread, cheese or bread=>butter.

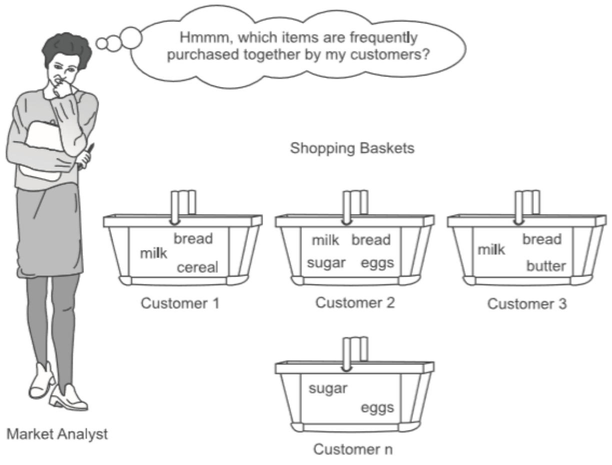


Fig-3: Need for association rule mining [4]

Fig-3 represents the need for (ARM) association rule mining. These rules are crucial for market analysist to determine which items are purchased together by the people and how these rules help any company to increase their profit.

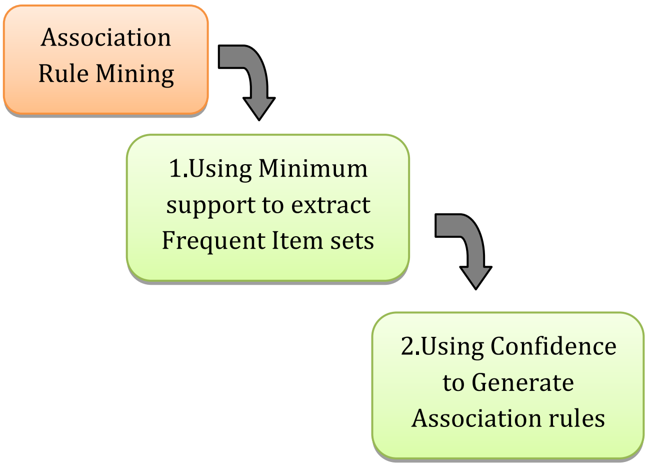

Association Rule Mining
1.Using Minimum support to extract Frequent Item sets
2.Using Confidence to Generate Association rules

Fig-4: Finding Association Rules

As shown in Fig-4, for finding association rules it involves two steps i.e. extracting frequent items from the dataset and then generating rules from these frequent items.

2.1 Measures in Association Rules

For any rule (X=>Y) to be strong and valuable it must satisfy the following measures [4]:

- **Support =** Frequency of X and Y occurring together divided by total transactions.

  Support =(Frequency of (X∪Y))/(Total Transactions)

- **Confidence =** Ratio of Frequency of X and Y occurring to the Frequency of X.

  Confidence = (Frequency of (X∪Y))/ (Frequency of (X))

- **Lift =** Confidence of X→Y divided by the frequency of Y

  Lift = (Confidence of (X→Y))/ (Frequency of (Y))

2.2 Association Rule Mining Algorithms

Different kinds of Patterns and algorithms in ARM are:

Frequent pattern: Items that will be occurring frequently in a data set. Example: Pants and shirts are bought together in a store. Algorithm like Frequent Pattern (FP)-Max, Equivalence Class Transformation (Eclat) etc. used for mining of the frequent Item sets.

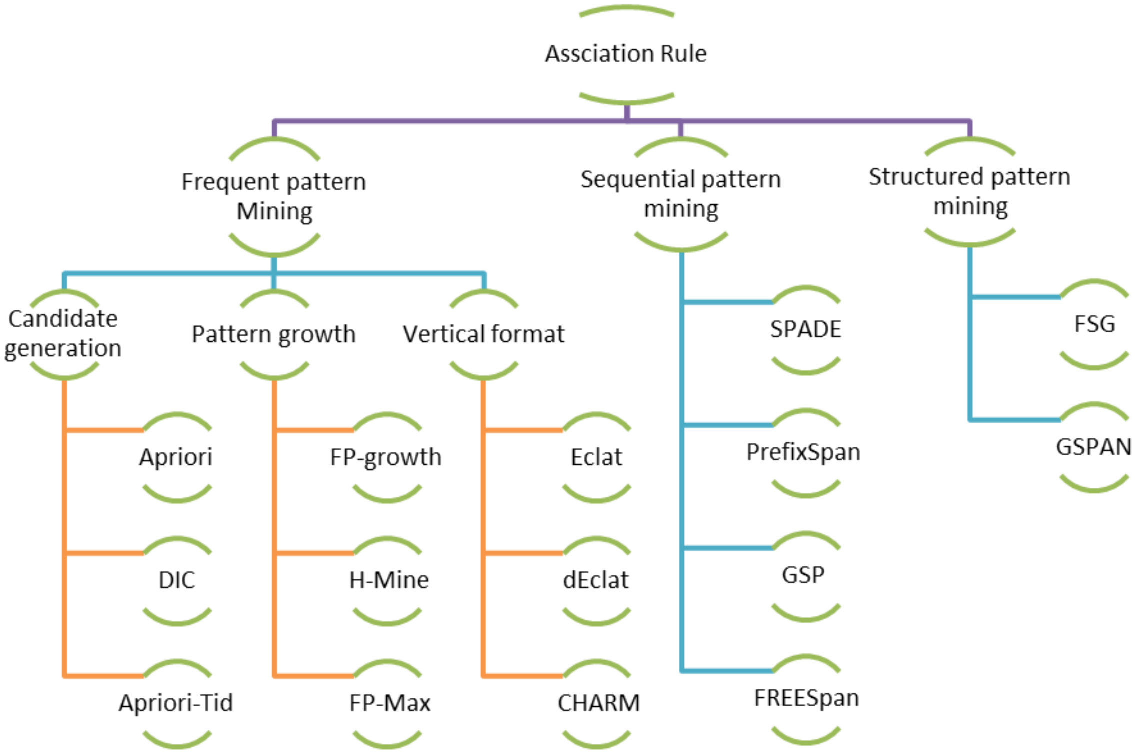


Fig-5: Different patterns and algorithms in ARM

Frequent pattern algorithms can be divided into three main types as shown in Fig-5 which are [1]:

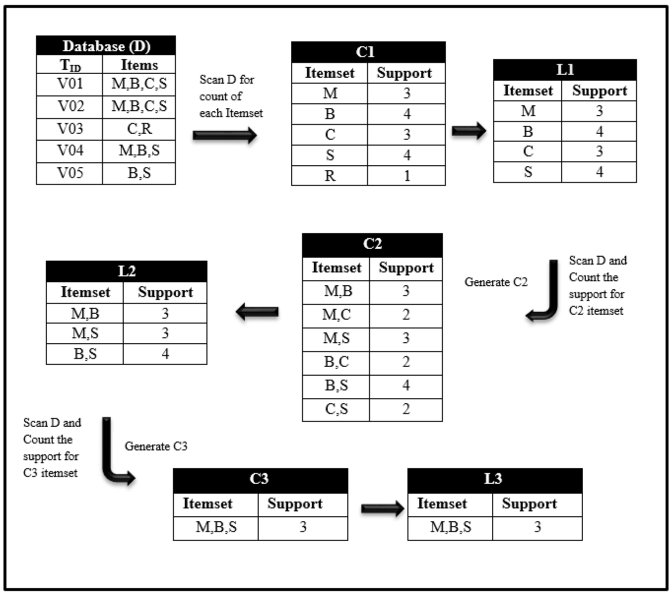
- **Candidate generation:** In these types of algorithms candidate set generation will take place for uncovering frequent itemsets and these algorithms are also time consuming. Apriori, (DIC) Dynamic Itemset Counting, Apriori-Tid (Transaction id), partitioning, sampling are some of the popular candidate generation algorithms.
- **Pattern growth:** In pattern growth algorithms, FP-growth is the most popular algorithm where mining of frequent itemsets take place without candidate generation. Some others pattern growth algorithms are H-Mine, FP-Max etc.
- **Vertical format:** In this, layout of the dataset is converted from horizontal format into the vertical format before applying the algorithms. Some of the most popular algorithms that works on the vertical layout of the database are Eclat, diffset Eclat (dEclat), Closed Association Rule Mining (CHARM) etc.

Sequential pattern: The items that are frequently purchased in a particular sequence. Example: The customer will first buy Laptop followed by the antivirus software then pen drive then keyboard. Algorithm like FREESpan, generalized sequential patterns (GSP) etc. used for sequential pattern.

Structured pattern: Structured pattern like trees etc., which are used for text retrieval and web analysis etc. Algorithm like Graph-based substructure pattern mining (GSPAN) etc. used for mining of the structures patterns [6].

Apriori Algorithm

Apriori is one of the prominent and extensively used



Database (D)	
T_{ID}	Items
V01	M,B,C,S
V02	M,B,C,S
V03	C,R
V04	M,B,S
V05	B,S

Scan D for count of each Itemset

C1	
Itemset	Support
M	3
B	4
C	3
S	4
R	1

L1	
Itemset	Support
M	3
B	4
C	3
S	4

Scan D and Count the support for C2 itemset
Generate C2

C2	
Itemset	Support
M,B	3
M,C	2
M,S	3
B,C	2
B,S	4
C,S	2

L2	
Itemset	Support
M,B	3
M,S	3
B,S	4

Scan D and Count the support for C3 itemset
Generate C3

C3	
Itemset	Support
M,B,S	3

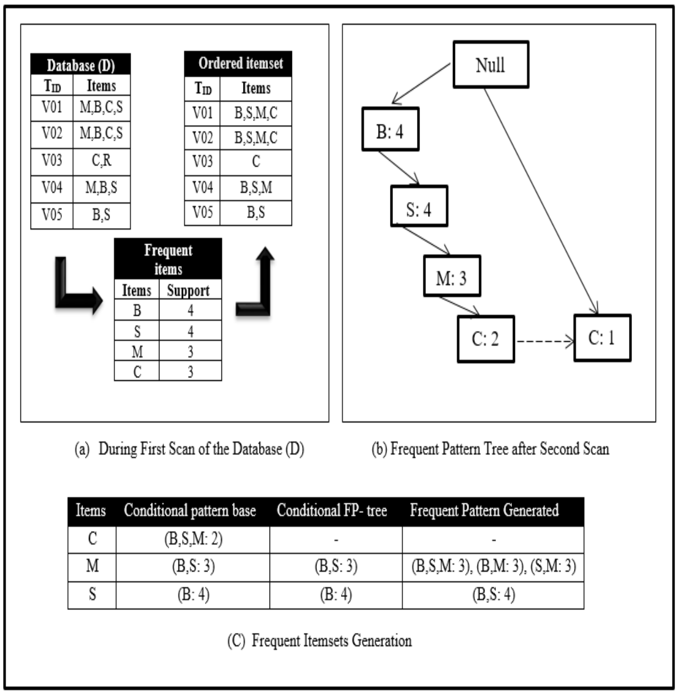
L3	
Itemset	Support
M,B,S	3

Fig-6: Example of Apriori Algorithm

algorithm for finding association rules and frequent itemsets to discover pattern from the database. It works on the horizontal layout database and also scan it whenever a new candidate itemset is generated. The drawbacks of the Apriori algorithm is that it generates multiple candidate itemset and scans the original database multiple time to find the support count of items [5]. Example: Generation of candidate itemset(C) and frequent itemset (L), where minimum support is 60% (3 transactions) in Fig-6. In Fig-6 first D is scanned to get the support of items and delete the items that does not satisfy minimum support condition to get L1, then generate C2 from L1 by performing self-join operation. Again delete items from C2 and generate C3 from L2 and at last L3 from C3, the final result are M, B, S it means M, B, S are frequently occurred.

FP-Growth Algorithm

Mining of frequent items faster and without candidate generation is achieved with the help of FP-Growth Algorithm. By scanning the dataset only twice i.e. one for getting ordered item sets and another to create FP tree. Also covers the weak points of Apriori algorithm [5]. Example: Generation of frequent itemset and FP-tree where the minimum support count is 60% (3 transactions) in Fig-7. As shown in Fig- 7 in (a) during first scan of D, first frequent items are found and then arrange the items of original database to get ordered itemsets in decreasing order of support.



Database (D)	
T_{ID}	Items
V01	M,B,C,S
V02	M,B,C,S
V03	C,R
V04	M,B,S
V05	B,S

Frequent items	
Items	Support
B	4
S	4
M	3
C	3

Ordered itemset	
T_{ID}	Items
V01	B,S,M,C
V02	B,S,M,C
V03	C
V04	B,S,M
V05	B,S

(a) During First Scan of the Database (D)

(b) Frequent Pattern Tree after Second Scan

Items	Conditional pattern base	Conditional FP- tree	Frequent Pattern Generated
C	(B,S,M: 2)	-	-
M	(B,S: 3)	(B,S: 3)	(B,S,M: 3), (B,M: 3), (S,M: 3)
S	(B: 4)	(B: 4)	(B,S: 4)

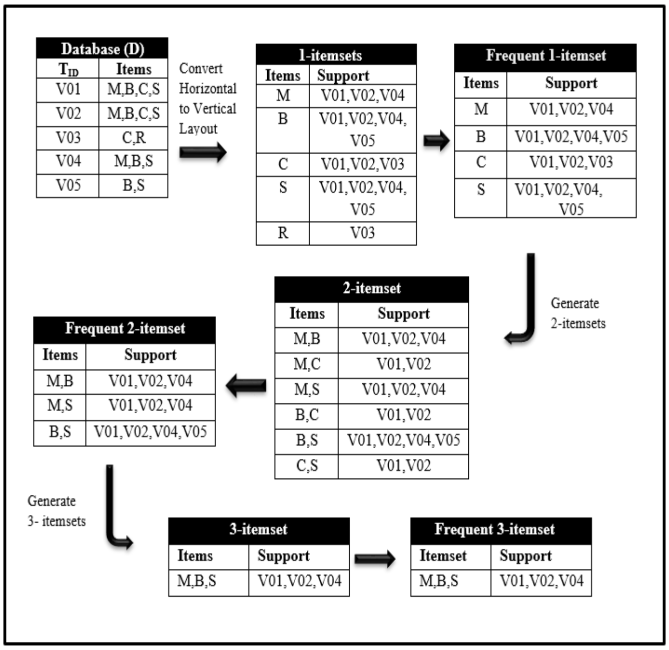
(C) Frequent Itemsets Generation

Fig-7: Example of FP-Growth Algorithm

In (b) again scan the database to construct FP-Tree and then in (C) Generating Frequent Patterns using Conditional pattern base and FP – tree. The minimum support count is 60% i.e. 3 transactions so frequent pattern with three or more transactions will only be considered such as (B, S, M: 3), (B, M: 3), (S, M: 3) and (B, S: 4).

Eclat Algorithm

The main job of the Equivalence Class Transformation algorithm is to find the frequent items in the group of transactions and it works only on the vertical layout database.



Database (D)	
T_{ID}	Items
V01	M,B,C,S
V02	M,B,C,S
V03	C,R
V04	M,B,S
V05	B,S

1-itemsets	
Items	Support
M	V01,V02,V04
B	V01,V02,V04, V05
C	V01,V02,V03
S	V01,V02,V04, V05
R	V03

Frequent 1-itemset	
Items	Support
M	V01,V02,V04
B	V01,V02,V04,V05
C	V01,V02,V03
S	V01,V02,V04, V05

2-itemset	
Items	Support
M,B	V01,V02,V04
M,C	V01,V02
M,S	V01,V02,V04
B,C	V01,V02
B,S	V01,V02,V04,V05
C,S	V01,V02

Frequent 2-itemset	
Items	Support
M,B	V01,V02,V04
M,S	V01,V02,V04
B,S	V01,V02,V04,V05

3-itemset	
Items	Support
M,B,S	V01,V02,V04

Frequent 3-itemset	
Itemset	Support
M,B,S	V01,V02,V04

Fig-8: Example of Eclat Algorithm

In this algorithm only support is calculated and confidence is not calculated. In this algorithm by intersecting the transaction id list of the subsets, support of an item set is calculated [6]. The main limitation of this algorithm is that it take more memory and time in case of very large databases because they contain very large transaction id sets [5]. Example: Generation of itemset and frequent itemset, where the minimum support count is 60% (3 transactions) in Fig-8. In Fig-8 first D is scanned to Convert Horizontal to Vertical Layout and delete the items that does not satisfy minimum support condition to get frequent 1-itemsets, then generate 2-itemset from frequent 1-itemset by using intersection method. Again delete items from 2-itemset and generate 3-itemset from frequent 2-itemset and at last frequent 3-itemset from 3-itemset, the final result are M, B, S it means M, B, S are frequently occurred.

3. LITERATURE REVIEW

M. Sinthuja et al. [7] presented an appraisal of association rule algorithms. They evaluated the performance of FP-Growth, Eclat and Apriori algorithms based on runtime and memory with respect to different support values and conclude that the performance of FP-Growth is better than other in terms of time and usage of memory. Kanwal Garg et al. [8] analyzed various algorithms used for frequent pattern mining namely FP-Growth, Eclat and Apriori. By differing the number of instances and attributes on synthetic datasets the performance was compared and parameter taken was execution time. The results obtained is, that FP- Growth is superior to other two algorithms. Subhash Rohit [5] presented survey on ARM algorithms and also gives their comparative analysis on various parameter like time, technique, data format and many others. Pinar Yazgana et al. [6] presented literature survey on ARM algorithms in which they group them on the basis of structured, Frequent and sequential pattern. Bhabesh Nath et al. [9] presented a comparative scrutiny of pattern mining algorithms and techniques namely frequent, rare, closed frequent and maximal frequent. The study infers that dataset characteristics as well as others factors like average transaction and pattern length greatly affects the performance of algorithm. Thabet Slimani et al. [11] presented a brief review on ARM. They have discussed about different kinds of frequent patterns and approaches used of pattern mining. M.Supriyamenon et al. [12] presented a review on association rule hiding techniques. They also presented a comparative analysis of ARM algorithms based on theoretical considerations and runtime. K. Vani [13] presented a comparative study based on performance survey in which Apriori performs worst and then experimentally evaluates the Eclat and FP-growth algorithms based on runtime. The paper concludes that increasing the support values will decrease the execution time of the algorithms. Trupti A. Kumbhare et al. [15] presented an overview of ARM algorithms namely AIS, Apriori, FP-growth and compare them on different characteristics like speed, data support etc. They conclude that FP-growth performs better than Apriori and AIS. Komal Khurana et al. [16] presented a comparative analysis on ARM algorithms based on accuracy, speed and data support. They conclude that Apriori Hybrid performs better compared to Apriori-Tid and Apriori. Jagmeet Kaur et al. [17] presented a survey on ARM in which they summarize the different applications of ARM and also gives an overview of different algorithms like genetic, Apriori and FP-growth algorithms.

4. RESULTS AND ANALYSIS

The comparison of different pattern mining algorithms is shown in Table-1 which is done based on some relevant criteria. For small itemsets Eclat takes less memory and time than FP-growth and Apriori algorithms. FP-growth takes only two database scans and does not generate any candidate sets compared to Eclat and Apriori algorithms. The data support in Eclat is vertical while in FP-growth and Apriori it is in

Table-1: Comparison of Frequent Item set Mining algorithms

Algorithm ➡ **Criteria ⬇**	**FP-Growth**	**Eclat**	**Apriori**
Technique [5]	Divide and conquer	Depth first Search	Breath First search
Strategy [9]	Tree based	Level wise Search	Level wise Search
Database scan [5]	Database is scan two time only	Database is scan few time	Database is scan each time candidate set is generated
Data support [12]	Very Large	Large	Limited
Itemset Generated [9]	Frequent	Frequent	Frequent
Execution Time [5]	Execution time takes less compare to Apriori	Takes less time compare to FP- Growth and Apriori	Execution time is more.
Methodology [14]	Conditional frequent pattern tree	Intersection of Tids list to generate candidate item set	Join and Prune
Data format [5]	Horizontal	Vertical	Horizontal
Data source [11]	Transaction database	Transaction database	Transaction database
Memory utilization [13]	Due to compact structure require less memory	Require less memory as compare to Apriori if itemsets are small in number	Require large memory space
Storage Structure [5]	Tree	Array	Array
Accuracy [13]	More accurate	More accurate	Less
Merit [9]	Less data base scans	Fast support counting	Generates more frequent itemsets
Privacy preserving approach preferred [12]	Reconstruction based approach	Reconstruction based approach	Heuristic approach
Demerit [9]	FP-Tree for large data may not fit into memory	Huge memory consumption	Huge memory consumption
Application [13]	Used in cases of large data	Best used for free item sets	Best used for closed item sets.

horizontal layout. The itemset generated by all three algorithms are frequent items and takes single support value. Overall FP-Growth and Eclat performs better compared to Apriori algorithm.

5. CONCLUSION AND FUTURE SCOPE

Association rules and frequent items are one of the most important research areas in data mining. It can also be absorbed in our daily life such as in the market some people buy milk, bread together while some people buy dress, jacket together or bed sheets and pillow covers together. This paper gives a brief introduction of important concepts and measures of ARM. In this paper, three pattern mining algorithms are compared based on some relevant criteria like execution time, accuracy, database scan and usage of memory. This study concludes that FP-growth and Eclat outplays the Apriori algorithm in terms of execution time and usage of memory. In Future the Apriori algorithm can be improved of by using machine learning techniques.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jain Pei, “Data Mining Concept and Technique”, 3rd Edition.

[2] Pang-Ning-Tan, Michael Steinbach, Vipin Kumar, “Introduction to Data Mining”, Pearson Publication, 2009.

[3] https://www.statisticshowto.com/data-mining, Accessed on : 3/08/2021 at 12:25 P.M.

[4] Parteek Bhatia “Data Mining and Data Warehousing: Principles and Practical Techniques”, Cambridge University Press, 2019.

[5] Subhash Rohit, “Association Rule Mining Algorithms: Survey”, International Research Journal of Engineering and Technology, Volume: 03, Issue:10, Oct-2016.

[6] Pinar Yazgana and Ali Osman Kusakci, “A Literature Survey on Association Rule Mining Algorithms”, Southeast Europe Journal of Soft Computing, Vol.5 No.1 March 2016.

[7] M.Sinthuja, N. Puviarasan and P. Aruna, “Evaluating the Performance of Association Rule Mining Algorithms”, World Applied Sciences Journal, Volume 35, Issue 1, 43-53, 2017.

[8] Kanwal Garg and Deepak Kumar, “Comparing the Performance of Frequent Pattern Mining Algorithms”, International Journal of Computer Applications, Volume 69, Issue 25, 21-28, May 2013.

[9] Anindita Borah and Bhabesh Nath, “Comparative evaluation of pattern mining techniques: an empirical study”, Complex & Intelligent Systems, volume 7, Issue 2, 589-619 (2021).

[10] Margaret H. Dunham, “Data Mining: Introductory and Advanced Topics”, 2006

[11] Thabet Slimani and Amor Lazzez, "Efficient Analysis of Pattern and Association Rule Mining Approaches", International Journal of Information Technology and Computer Science, vol.6, no.3, pp.70-81, 2014.

[12] M.Supriyamenon and P.Rajarajeswari, “A Review on Association Rule Mining Techniques with Respect to their Privacy Preserving Capabilities”, International Journal of Applied Engineering Research, Volume 12, Number 24 (2017).

[13] K. Vani, “Comparative Analysis of Association Rule Mining Algorithms based on Performance Survey”, International Journal of Computer Science and Information Technologies, Vol. 6 (4), 2015, 3980-3985

[14] Vikram Rajpoot, Shanu kumar and Sadhna K. Mishra, “Review On Frequent Itemset Mining Algorithms in Data Mining”, International Journal of Technical Innovation in Modern Engineering & Science, Volume 4, Issue 09, September-2018.

[15] Trupti A. Kumbhare and Santosh V. Chobe, “An Overview of Association Rule Mining Algorithms”, International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, 927-930.

[16] Komal Khurana and Simple Sharma, “A Comparative Analysis of Association Rules Mining Algorithms”, International Journal of Scientific and Research Publications, Volume 3, Issue 5, May 2013.

[17] Jagmeet Kaur and Neena Madan, “Association Rule Mining: A Survey”, International Journal of Hybrid Information Technology, Vol.8, No.7 (2015), pp.239-242.