A Comparative Analysis of Association Rule Mining Algorithms

Akash Saxena¹ and Vikram Rajpoot²

¹Professor & Head, Department of CSE, CITM College, Jaipur, Rajasthan, India
²Assistant Professor, Department of CEA, GLA University, Mathura, UP, India

E-mail: akash27jaipur@gmail.com

Abstract. The field of data mining (DM) has grown rapidly in recent years. One of the most important data mining techniques is association rule mining (ARM). It is a strategy used to identify trends in the database that are normal. There has been a lot of work in the area of ARM. The paper provides a short description of the principles and algorithms of interaction, several of the implementations. To several researchers, ARM has long been and still is of concern. Data mining is one of the essential activities. This helps to identify associations between various elements in the database. The goal of this paper is to provide an outline of the fundamental concepts of the ARM methodology and the recent relevant research in this area. The paper further explains the different algorithms, methods, strategies, and benefits of the ARM areas, drawbacks. The paper also provides a minor distinction focused on the results of various algorithms related to association rules mining. The paper provides a short description of the principles and algorithms of interaction, several of the implementations. Algorithms are present and evaluate base parameters such as precision, algorithm pace, and help for data. To solve the question of apriori algorithms AprioriTid and the AprioriHybrid have been suggested. From the contrast, we infer that, since it has decreased overall pace and increased precision, AprioriHybrid is superior to Apriori and AprioriTid. We may infer that the LogElcat algorithm performs more than every other algorithm based on these parameters.

1. Introduction

Data mining is a method by which valuable data are obtained from the databases. It is designed to manage vast volumes of data. Future trends and behavior are predicted by DM. DM is becoming a useful technique in these days because it can be used in different areas. Across various fields, a tremendous amount of data is generated per day. A vast volume of data cannot be analyzed and data processing is often used. It is also called knowledge discovery in database (KDD). Tools for DM are used to predict the data pattern & behavior so that organizations can make their decisions [1]. Because of its popularity in various fields, data mining is popular. It extended itself in fields like healthcare, banking, telecommunications, business, schooling, and several others [2]. Two types of data mining are concise and prescriptive. We identify & summarize the existing information in descriptive mining. Historical details are used for estimation in prescriptive mining [3]. We have various strategies in data mining, which include classification, clustering, ARM.

One of the most important data mining techniques is association rule mining. This is used to identify the trends in the sample that are more common. Their main purpose is to define the interesting connections
among the different elements of the database [3]. The ARM is focused on knowledge. For the identification of data trends, correlation principles are used in several areas. Using trends, we know how many variations of incidents arise concurrently. On large amounts of data, Association rules are implemented. AR is used in business to discover the common trends which contribute to marketing and decision-making. Many other areas still have association rules which involve medical research, business basket analyzes, library research, etc. [4]. For instance, if you buy a shampoo, you can buy a conditioner as well. These kinds of relationships can be used as information that is useful in making marketing decisions [5].

2. Overview of Data Mining
DM is a discipline at the convergence between computer science & analytics and is often called information exploration on datasets. The method seeks to uncover the trends between vast datasets. It uses artificial intelligence, ML, statistics & database systems intersection techniques. The overall objective of the DM process is to extract information from a data set & transform it into a comprehensible structure. Data mining's key roles are to use diverse techniques and algorithms to find and remove patterns of data collected [2]. Since its value in decision-making, data mining, or information creation technologies have been an important feature of numerous organizations. In addition to statistics, database, machine learning, pattern reconstruction, artificial intelligence, computing skills, and other areas, data mining strategies have been implemented [6].

2.1. Data Mining Process
KDD process (knowledge discovery in a database) is also known as the DM process. Extraction or exploration of information is carried out in seven phases following DM:

a. Data cleaning: We extract noise & incompatible details from the raw information in this phase.
b. Data integration: Different data are here merged into one data for data purposes.
c. Data Selection: We will then access the task-related data that we need more.
d. Data Transformation: Where data is converted via a description or aggregation process into suitable mining types.

![Figure 1. Data Mining Process](image_url)
e. **Data Mining:** Different methods or resources for collecting information are used here.
f. **Pattern evaluation:** In this stage, the trends of DM values are established.
g. **Knowledge representation:** In this step, visualization and representation of knowledge are done to aid users in understanding mined knowledge [7][8].

2.2. **Data Mining Techniques**
Various major DT techniques, like affiliation, rule classification, clustering, prediction as well as measurement trends, etc, have been residential and applied to DT projects recently and are used in database exploration.

2.2.1. **Association**
It is a popular technique in DM. Throughout this methodology; we evaluate recurrent trends and identify interesting ties inside results.

2.2.2. **Classification**
The role of searching for a pattern or function defines and differentiates data groups or concepts to determine where an entity is placed whose class mark is unknown.

2.2.3. **Clustering**
Clustering is a method of classifying a collection of specific or abstract items of related artifacts. A cluster is a group of artifacts that are "related" and "different" from the artifacts in other clusters. A cluster is the set of data points that are comparable and separate from artifacts in other clusters.

2.2.4. **Prediction**
The classification predicts categorical names, prediction models, which are continuously priced. This is to estimate numerical data values lost or inaccessible instead of class labels. Nevertheless, both the numerical and class mark predictions can often apply to word estimation [9].

3. **Association Rule Mining**
Association rule mining (ARM) is used to define the relation between a large numbers of data objects. Thanks to the vast volume of data in repositories, many companies are concerned about their database mining association rules as shown in figure 2. For example, catalog design, cross-marketing, and other business decision-making processes can be assisted by the recognition of the fascinating related relationships between large volumes of transaction data. Market basket analysis is a typical example of ARM. This method explores consumer purchasing habits by detecting correlations between various products put by customers in their packs. The detection of such partnerships will enable marketers to broaden communication campaigns by providing an overview of the products they mostly buy together. The researchers will create an improved data mining algorithm in this study [10].

![Generating Association Rules](image-url)

---

**Figure 2. Generating Association Rules**
3.1. APRIORI
Apriori is an ARM technique that removes all frequently occurring items from the transaction when inputting transactional databases. In this case, provided the electronic medical record in Apriori, a variety of commonly occurring risk factors are developed and certain factors for the development of diabetes are indicated.

3.2. ELCAT
Elcat is similar to the recursive operation of the Apriori algorithm. It uses a tree such as the Tidset framework. In turn, the EMR produces the common risk factors for diabetes. The tidset starts with all risks in the database once in a while. The algorithm performs an in-depth search to assess the risk factor in all situations.

3.3. FP- Growth Algorithm
Used to resolve the drawbacks of the program apriori. The document just has to be searched two times here. There is a need to create an FP tree in this algorithm, so you use two passes.

Pass 1:
- The index is searched such that help and frequent1 objects are identified.
- The rare items are thrown out.
- The things are in the lower service order.

Pass 2:
- Many items with higher funding would be found. FP- tree is used for removing can objects.

4. Genetic Algorithm
Throughout the fields of computation, biological verification, and machine learning the method is practically applied. This generates new population cords using the old cords in an iteration way. The strings are binary encoded, true, and so on. Genetic operations, range, mutation, and interposition, are implemented in this algorithm [11][12].

Table 1. Some ARM Algorithms of Advantages & Disadvantages

| ARM Algorithm | Advantages | Disadvantages |
|---------------|------------|---------------|
| AIS           | 1. An estimate of these candidate items, which1. Only one object is subsequently restricted. have no great hope of being, is used in the2. Several passes are needed via the database. algorithm. 2. The incomplete contract database may be used for small specified data structures required for cardinality purposes. 3. There are no the maintenance of large and candidate items | |
5. Applications of Association Rule Mining

ARM is used in different fields because it is highly helpful. Some of the areas protected by association rule mining are discussed below.

5.1. Market Basket Analysis

Market basket research, for example at grocery shops, is one of the strongest and common forms of related rule mining. Managers in these businesses aim to draw potential consumers and raise the desires of their existing customers. There are no such stores, and an enormous number of transaction records are present.

5.2. CRM of the Credit Card Business

Customer Relationship Management is a method that a bank uses to monitor customer relations. Banks typically classify their customers' actions to uncover their tastes and desires. This will enhance the coherence between bank card clients.

5.3. Medical Diagnosis

In the field of medical sciences, association rules are also applicable. The patients will be treated by surgeons. To determine the probability of illness, association rules are used. An appropriate reason for a disease is important.

5.4. Census Data

Census data have great potential for association rule mining. Association rules aim to encourage effective public policy as well as the growth of enterprises. A census produces tremendous statistical knowledge.

5.5. Protein Sequences

The essential components of every organism's cell are proteins. Many DNA technologies with various tools for the speedy determination of DNA sequences are available. Proteins are the sequences composed of almost 20 amino acid forms. The structure of a protein depends on the amino acid sequence [12].

6. Literature Survey

Rahman, T., (2019) Interesting rules are generated in the field of data mining based on multiple criteria for the classification of objects. Associations rule mining is used using the following three techniques-
conditional, quantitative, and Fuzzy association rule mining for the generation of rules dependent on help, trust, boost, leverage, etc. The Fuzzy association rule mining uses linguistic values as well as member functions to determine the rules so that the other two association regimes mining techniques provide more realistic values. The Fuzzy logic system, with an evolutionary algorithm architecture, can overcome this lack rather than using flat and measurable values for rule mining. The performance of the Fuzzy association rules algorithms for mining is compared in this paper. The classic Apriori, Fuzzy-Apriori, and Evolutionary Genetic Fuzzy Apriori DC algorithms are used to analyze performance for performance [14]. Naresh, P., & Suguna, R. (2019) This study compares main ARM algorithms including Apriori, FP-Growth, LCM, or FIN. The findings are extracted, analyzed, and compared with both of these methods with large databases and limited data sets. FIN is the newest algorithm from which frequency sets are created quickly. LCM is an algorithm that received the prize. FP-Growth generates frequent itemsets without the Apriori approach of candidate generation. These algorithms are selected for comparative analysis because of their major differences. This article provides descriptions of the algorithms and diagrams, as well as contrasts in terms of output metrics such as execution time and quality [15]. Pornthep Rojanavasu (2019) This paper suggests the usage of two educational data mining techniques. First, in admission data association rules are introduced to identify other information to facilitate admission preparation. Secondly, the decision-making tree was used to forecast employment after they graduate grades and work details. The research findings offer strong insights into intake preparation and work forecasting [16].

Guihong Zhang (2019) The current data mining technology has inconsistency of using keywords & words, resulting in low mining accuracy and efficiency. A new DM algorithm is proposed in this article based on an AR algorithm. The K-means clustering algorithm conducts the clustering study of new mining results, and the correspondence for the resulting mining results is calculated according to the decreased order. The test findings show an accurate and efficient algorithm [17].

Cao, M., & Guo, C. (2017). The power monitoring device collects large quantities of data used for alarm purposes only. This information contains a wealth of knowledge but has not been excavated effectively. so, an advanced interaction law algorithm is implemented in this paper based on Upper Triangular Matrix (UTM). The algorithm allows the size of candidate sets to be streamlined and for the database to be reduced to a minimum. There is thus a significant improvement in the efficiency of rules acquisition. An experiment is done through an association rule analysis to verify the feasibility & validity of the improved algorithm in power monitoring data. It provides an important path to data extraction and the exploration of association rules [18].

Soni, H. K., (2017) Within this paper we carried out a detailed analysis of different association rules on e-commerce algorithms. We also identify the weaknesses of several existing algorithms. There are also certain plausible characteristics suggested for the design of an efficient e-commerce database algorithm to support incremental as well as interactive ARM [19].

Patil, A., & Gupta, P. (2017). In the area of data mining, ARM plays a significant position. The mining rule of Association is a technique for preparing ways to improve mining technology. Various types of approaches or algorithms have been established in the combination rule mining but it is very necessary to know the most appropriate approach to the combination rule. In this article, we introduce a new algorithm named upgrowth that has the potential to reduce the complexity of time with the high dimensional details [20].

Bon-Hong Koo, (2016) In this respect, a mining concept, the Terrorist Network Mining, is the solution for data mining. The guerrilla mining network has proved to be a workable solution for terrorist detection and analysis. This concept still had to be improved by connecting fuzzy with genetic processes with IDS, leading to a substantial and efficient detection process. Therefore, there were still some improvements needed. The paper explores how well in tandem with useless data mining (reveal trends whose action is intrusive) an intrusion detection method does following the genetic algorithm (conducting effective intruders detection) [21].
7. Comparative Analysis

The analysis is as shown in table 2 to 4 and in figures 3 and 4.

| Features                  | AIS | SET M | Apriori | Apriori TID | Apriori hybrid | FP-growth |
|---------------------------|-----|-------|---------|-------------|----------------|-----------|
| Data support              | Less| Less  | Medium  | Often large | Very Large     | Very Large |
| Speed in the initial phase| Slow| Slow  | High    | Slow        | High           | High      |
| Speed in the later phase  | Slow| Slow  | Slow    | High        | High           | High      |
| Accuracy                  | Very less | Less  | Less    | Medium, More accurate than Apriori | Fast, More accurate than Apriori TID | Very fast, More Accurate |

Table 3. Comparisons Representation of the Performance of ARM Algorithm [23]

| Features                  | AIS | SET M | Apriori | Apriori TID | Apriori hybrid | FP-growth |
|---------------------------|-----|-------|---------|-------------|----------------|-----------|
| Data support              | 1.5 | 1.5   | 2.5     | 4           | 5              | 5         |
| Speed in initial phase    | 1.57| 1.57  | 4.98    | 1.3         | 5              | 5         |
| Speed in later phase      | 1.58| 1.58  | 1.5     | 5           | 5              | 5         |
| Accuracy                  | 0.5 | 0.5   | 1.45    | 2.46        | 4              | 5         |
Figure 3. Graphical Representation of Association Rule Mining Algorithms

Table 4. Comparisons Representation of the Execution Time (in ms) of ARM Algorithm

| Algorithm    | Apriori | AprioriTID | Apriori hybrid | FP-growth | Eclat |
|--------------|---------|------------|----------------|-----------|-------|
|              | 232     | 160        | 105            | 60        | 30    |

Figure 4. Time to Execute the Algorithm

As per figure 4, Apriori takes longer to execute the algorithm and does not function well in big datasets, whereas Fp-Growth is hindered by its complex trees which increases the algorithm’s difficulty.

8. Conclusion
The rule of association mining among marketers has become especially common. Indeed, an example is called a market basket analysis of the ARM. The role is to find out which objects are always gathered.
Professionals may use this information to prepare concepts and position items frequently bought near each other, thus maximizing sales. The law of association mining involves the connections between objects in a data collection. The algorithm of ARM is discussed in this paper. The most appropriate ones are contrasted from a broad variety of effective algorithms. This also provides a comparative analysis of different strategies for ARM which algorithm will be a better fit. This paper provides a study of the algorithms of ARM. Algorithms AIS, SETM, AprioriTid, AprioriHybrid, FP-Growth as well as LogEclate. Comparison is focused on factors such as accuracy, algorithm speed, support of the database, etc. Every algorithm takes more time to execute the algorithm with a few benefits and drawbacks, while its complex structure is disadvantageous to Fp-Growth that increases the complexity of an algorithm. From this basis, we can infer that LogEclat is a better algorithm than any other.

9. References

[1] N. Padhy, P. Mishra and R. Panigrahi, “The Survey of Data mining Applications and Feature Scope”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), vol. 2, no. 3, (2012) June.

[2] A. N. Paidi, “Data Mining: Future Trends and its Applications”, International Journal of Modern Engineering Research (IJMER), vol. 2, no. 6, November-December (2012), pp. 4657-4663.

[3] T. Karthikeyan and N. Ravikumar, “A Survey on Association Rule Mining”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 1, (2014) January.

[4] A. Rajak and M. K. Gupta, “Association Rule Mining: Applications in Various Areas”, International Conference on Data Management.

[5] Arora, Amandeep Singh, Linesh Raja, and Barkha Bahl. "Data centric security approach: A way to achieve security & privacy in cloud computing." Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT). 2018

[6] M. Manohara, R. Dinesh, and S. M. Sowmya,” Interactive Data Mining: A Brief Survey”, Lecture Notes on Software Engineering, Vol. 1, No. 2, May 2013

[7] Nikita Jain,” Data Mining Techniques: A Survey Paper”, IJRET: International Journal of Research in Engineering and Technology elISSN: 2319-1163 | pISSN: 2321-7308, Volume: 02 Issue: 11 | Nov-2013.

[8] Jain, Nikita, and Vishal Srivastava. "Data mining techniques: a survey paper." IJRET: International Journal of Research in Engineering and Technology 2.11 (2013): 2319-1163.

[9] Md Zahidul Islam,” Privacy Preservation in Data Mining Through Noise Addition”, November 2007. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.726.6772&rep=rep1&type=pdf

[10] Savi Gupta and Roopal Mamorta,” A Survey on Association Rule Mining in Market Basket Analysis”, International Journal of Information and Computation Technology., Volume 4, Number 4 (2014), pp. 409-414

[11] J.OMANA et al, International Journal of Computer Science and Mobile Computing, Vol.6 Issue.4, April- 2017, pg. 5-8.

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

[13] Gurneet Kaur et al, / (IICSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, 2320-2324

[14] Rahman, T., Kabir, M. M. J., & Kabir, M. (2019). Performance Evaluation of Fuzzy Association Rule Mining Algorithms. 2019 4th International Conference on Electrical Information and Communication Technology.

[15] Gupta, Shaurya, et al. "Tier application in multi-cloud databases to improve security and service availability." Handbook of research on cloud computing and Big Data applications in IoT. IGI Global, 2019. 82-93.
[16] Naresh, P., & Suguna, R. (2019). Association Rule Mining Algorithms on Large and Small Datasets: A Comparative Study. 2019 International Conference on Intelligent Computing and Control Systems (ICCS). DOI:10.1109/iccs45141.2019.9065836

[17] Pornthep Rojanavasu, ”Educational Data Analytics using Association Rule Mining and Classification”, and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering, 978-1-5386-8072-8/19/$31.00 ©2019 IEEE

[18] Guihong Zhang, (2019)” Research on Data Mining Technology based on Association Rules Algorithm”, 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAI2019).

[19] Cao, M., & Guo, C. (2017). Research on the Improvement of Association Rule Algorithm for Power Monitoring Data Mining. 2017 10th International Symposium on Computational Intelligence and Design (ISCID). DOI:10.1109/iscid.2017.72

[20] Soni, H. K., Sharma, S., & Jain, M. (2017). Plausible characteristics of association rule mining algorithms for e-commerce. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication, and Bio-Informatics (AEEICB). DOI:10.1109/aeeicb.2017.7972379

[21] Patil, A., & Gupta, P. (2017). A review of the up-growth algorithm using association rule mining. 2017 International Conference on Computing Methodologies and Communication (ICCMC). DOI:10.1109/icccmc.2017.8282605

[22] Bon-Bong Koo,” association rule mining and genetic algorithm (ga) for data mining based intrusion detection system: a review approach”, IEEE Journal on selected areas in communications, vol. 34, no. 3, March 2016.

[23] P. Prithiviraj and R. Porkodi, “A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study”, American Journal of Computer Science and Engineering Survey (AJCSES), vol. 3, issue 1, 2015, pp. 098-119.