Optimizing Data Exploration by Unifying Clustering and Association Rule Extraction

Youssef Fakir, Salim Khalil, Hamid Grmani, Mohamed Fakir
Department of Computer Sciences, Faculty of Sciences and Technics, Sultan Moulay Slimane University, Beni Mellal, Morocco
E-mail: m.fakir@usms.ma

Keywords: association rules, apriori, FP-Growth, k-means, clustering

Received: March 31, 2024

The extraction of association rules remains a crucial strategy in data analysis, particularly in the context of massive datasets. This method unveils complex relationships, correlations, and meaningful patterns within vast datasets, providing essential insights for decision-making and understanding behaviors. Our approach stands out through the use of clustering algorithms for intelligent data partitioning. This strategic choice establishes a robust foundation for efficient association rule extraction. By organizing data specifically through clustering techniques before applying the extraction algorithm, we aim to optimize the relevance and significance of the discovered rules.

Povzetek: Prispevek združuje K-means gručenje in algoritma Apriori/FP-Growth za optimizacijo odkrivanja asociacijskih pravil, s čimer izboljša učinkovitost analize velikih, heterogenih podatkovnih množic.

1 Introduction

In an era characterized by exponential data growth, organizations face the immense challenge of extracting meaningful insights from increasingly massive datasets. Data mining has emerged as an indispensable tool for uncovering valuable knowledge, with association rule extraction standing out as a fundamental strategy for identifying hidden relationships, unexpected correlations, and essential patterns within complex datasets [1]. This study focuses on an innovative methodology that integrates the K-means clustering algorithm for intelligent data partitioning, aiming to optimize the process of extracting association rules [2]. By efficiently segmenting data, this approach addresses challenges inherent to large datasets, enabling the discovery of highly relevant and actionable insights [3].

The primary objective of this work is to enhance the quality and applicability of extracted association rules, providing organizations with a targeted understanding of the intricate dynamics embedded in extensive datasets [4, 5]. Building on existing research, we explore the synergy between clustering and association rule extraction, leveraging K-means to create a solid foundation for robust data mining operations [6, 7]. Recent studies, such as those by [8,9] demonstrate that combining clustering techniques with rule extraction algorithms like Apriori and FP-Growth can significantly improve scalability and computational efficiency in big data analytics.

One of the most significant challenges in mining massive datasets lies in their heterogeneity and complexity. Traditional rule-mining approaches often struggle with issues such as high dimensionality, noise, and the computational cost of analyzing large-scale data [10]. Clustering algorithms like K-means mitigate these challenges by organizing data into meaningful groups, which can then be analyzed independently. This partitioning not only improves the efficiency of association rule extraction but also enhances the interpretability of results by focusing on specific data subsets, as demonstrated in applications like market basket analysis and healthcare analytics [11].

Furthermore, distributed computing frameworks, such as Apache Spark and Hadoop, have paved the way for implementing clustering and rule extraction algorithms at scale.

By distributing the workload across multiple nodes, these technologies enable the efficient processing of terabyte-scale datasets while maintaining high accuracy [12]. For example, recent advancements in distributed K-means clustering have been used in conjunction with Apriori to accelerate the discovery of meaningful patterns in e-commerce transaction data [13]. Our methodology builds upon these advancements, integrating distributed clustering techniques into a cohesive framework for association rule mining.

To illustrate the practical implications of our approach, we examine its application in real-world scenarios. One such example is the detection of fraudulent activities in financial transactions, where clustering algorithms help isolate anomalous patterns that indicate potential fraud. Similarly, in personalized marketing, the combination of K-means and FP-Growth has proven effective in identifying customer preferences and designing targeted campaigns [14]. These examples underscore the versatility and potential of our integrated approach, which leverages

the strengths of clustering and association rule extraction to address diverse challenges in big data analytics.

Our article is structured to provide a comprehensive understanding of our innovative approach to association rule extraction in a distributed environment. In section 2, we introduce the general context of massive data analysis and the importance of association rule extraction. Section 3 details our approach, highlighting the use of the K-means algorithm for intelligent data partitioning and the combination of association rule extraction algorithms such as Apriori and FP-Growth. We also present our methodology by describing the modeling steps used, while section 4 examines the obtained results and offers detailed analyses. Finally, section 5 concludes by emphasizing the importance of our integrated approach and future research prospects in this field. To further demonstrate the effectiveness and importance of clustering algorithms in data exploration, and to highlight the valuable insights gained from the implementation of these algorithms, we will now examine recent research in this area, including the learnings from their practical application.

2 State of the art

The state of the art in the field of association rule mining highlights the increasing importance of using clustering techniques to enhance the efficiency of this process. Recent works have emphasized the positive impact of integrating the K-means algorithm into association rule mining, especially for analyzing large datasets. For example, Li et al. proposed a closed-loop hierarchical clustering approach for optimizing unit commitment and dispatch in micro grids [4]. Similarly, Guha et al. [15] presented a robust clustering algorithm for categorical attributes, called ROCK [15], which demonstrated significant performance in data segmentation.

In comparison with these recent works, the presented article advocates for an innovative approach that synergistically combines clustering with association rule mining algorithms such as Apriori and FP-Growth. This approach aims to optimize the relevance and significance of discovered rules by preorganizing data using clustering techniques. By integrating these two aspects, the article offers a comprehensive methodology for association rule mining in a distributed environment, thus providing deeper insights into user behaviors and trends within massive datasets.

Furthermore, the article underscores the importance of understanding the unique characteristics of association rule mining algorithms such as Apriori and FP- Growth, and adapting the approach based on dataset specifics. This integrated approach highlights the effectiveness of data exploration through the combination of clustering and association rule mining techniques, offering promising prospects for more efficient and relevant data exploration practices.

3 Proposed approach

Presently, databases pose a significant challenge due to their extensive size and distribution across various locations [1]. This has resulted in the extraction of association rules from these databases becoming an intricate and resource-intensive task, marked by prolonged execution times and computational complexities [15].

Addressing these issues is imperative, and parallelism emerges as a crucial solution for efficient association rule extraction [6, 16]. Our exploration of sequential and parallel techniques revealed persistent challenges, notably the high volume of data scans and the detrimental impact of multiple synchronization and communication stages on algorithm performance [4, 5]. Recognizing these obstacles, we identified the sequential Partition algorithm as a potential remedy, requiring just two passes through the database [15]. To further alleviate the situation, our focus shifted towards enhancing the Partition algorithm to streamline the process to a single database scan, with subsequent consideration for the development of a parallel version [2]. Additionally, optimizing algorithm performance demands intelligent base partitioning using clustering, particularly distributed clustering [3]. In light of the comprehensive literature review presented in the preceding chapter, we closely examined sequential and parallel partitioning algorithms, recognizing them as promising avenues for achieving homogeneous database distribution across multiple sites [4, 5]. As a proactive step, we propose a parallel clustering algorithm based on the k-prototypes algorithm, deemed most effective for handling the complex nature of real-world data containing both numerical and categorical values [2]. The methodology approach proposed is illustrated in Figure 1.

3.1 FP-Growth algorithm

Algorithm 1: FP-Growth Algorithm

Input: Database of transactions, minimum support threshold minsup

Output: Frequent itemsets

1. **Construction of the FP-Tree**:
   Traverse transactions, count frequency of each unique item, remove infrequent items, sort based on frequency, and build FP-Tree.
2. **Construction of the header table:**
   Create a header table to record first occurrence of each item in the tree.
3. **Construction of conditional sets:**
   For each item in header table, extract conditional paths in the tree.
4. **Recursion to extract frequent sets:**
   Repeat process from Step 1 to extract frequent patterns from conditional sets.
5. **Generation of association rules:**
   Generate all possible combinations of items in frequent patterns, split into antecedent and

consequent, calculate confidence, and filter rules.

6. **Return the final set of association rules**: Final output is set of association rules meeting specified criteria.

3.2 K-means algorithm

Algorithm 2: K-means Algorithm

Input: dataset, number k

Output: Cluster centers, assignment of points to of clusters clusters

1 Randomly select k points as initial centers;
2 repeat
3 Assign each point to the cluster with the nearest center;
4 Recalculate centers by taking the mean of points in each cluster until convergence;
5 return Cluster centers, assignment of points to clusters

3.3 Apriori algorithm

Algorithm 3: Apriori Algorithm

Input: Database of transactions, minimum support (minsup)

Output: Frequent itemsets

1 Generate frequent 1-itemsets by scanning the database;
2 k ← 2;
3 Ck ← Candidate item sets of size k;
4 Lk−1 ← Frequent item sets of size k − 1;
5 while Lk−1 is not empty do
6 Ck ← Generate candidates from Lk−1;
7 Ck ← Prune candidates with infrequent subsets;
8 Count the support of each candidate in Ck by scanning the database;
9 Lk ← Keep candidates with support ≥ minsup;
10 k ← k + 1;
11 return Frequent itemsets

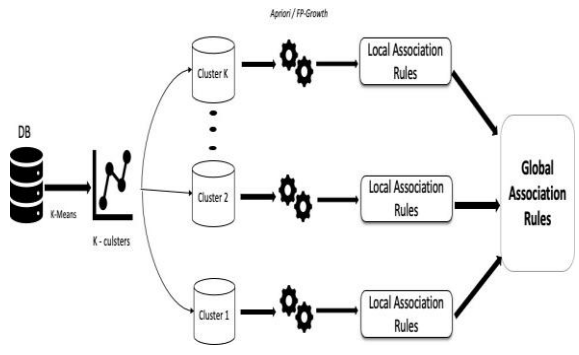

Figure 1: Methodology approach

3.4 Modeling with K-Means, apriori and FP-growth for association rule extraction

❖ **Step 1: Applying K-means for cluster creation**:

We begin our modeling by applying the K-Means algorithm on our original dataset [4]. The objective here is to group the data into separate clusters, each representing a homogeneous set of characteristics. After applying K-Means, each data point is assigned to a specific cluster based on its characteristics, creating a clear segmentation of our dataset [17, 18].

❖ **Step 2: Creating subsets of data:**

Once K-Means has created clusters, we divide the original dataset into separate subsets based on those clusters. Each subset contains the data associated with a particular cluster. This step is crucial to isolate the specific behaviors present in each homogeneous group.

❖ **Step 3: Apply the apriori algorithms, fp-growth to each subset:**

With our subsets in place, we now apply the Apriori and FP-Growth algorithms to each of them. The goal is to extract meaning association rules within each group. These algorithms analyze transactions in each subset, identifying specific buying or behavior trends for each cluster [14].

❖ **Step 4: Grouping the rules of association:**

Once we have generated association rules for each subset, we aggregate all these rules to obtain a consolidated view of the entire behaviors extracted from each cluster. This consolidation allows us to observe overall trends and correlations among different user groups.

Finally, it is essential to note that our study was carried out using the CC GENERAL dataset plays a crucial role in our research as it contains customer data related to credit card transactions. We used this dataset to analyze customer behavior, identify trends in spending patterns, and predict credit risk. It includes credit card data, customers, their balances, purchase and cash advance frequencies, and payment histories [19, 20]. Table 1 shows the columns and their descriptions of dataset.

Table 1: Description of dataset

Attributes	Description
CUST_ID	Unique identifier for each customer
BALANCE	Current balance of the customer's account
BALANCE_FREQUENCY	Frequency of balance updates (e.g., daily, weekly, monthly)
PURCHASES	Total purchases made by the customer.
ONEOFF_PURCHASES	Total one-off purchases made by the customer
INSTALLMENTS_PURCHASES	Total purchases made through installment plans
CASH_ADVANCE	Total cash advances taken by the customer
PURCHASES_FREQUENCY	Frequency of purchases (e.g., daily, weekly, monthly).
ONEOFF_PURCHASES_FREQUENCY	Frequency of one-off purchases
PURCHASES_INSTALLMENTS_FREQUECY	Frequency of purchases made through installment plans
CASH_ADVANCE_FREQUENCY	Frequency of cash advances.
CASH_ADVANCE_TRX	Number of cash advance transactions
PURCHASES_TRX	Number of purchase transactions
CREDIT_LIMIT	Credit limit of the customer's account
PAYMENTS	Total payments made by the customer
MINIMUM_PAYMENTS	Minimum payment amount required by the customer
PRC_FULL_PAYMENT	Percentage of customers who made full payments
TENURE	Tenure of the customer's account (e.g., years, months).

4 Results and analysis

In this section, we closely examine the performance of the Apriori algorithm in a distributed environment based on the number of clusters generated by our K-Means approach. The goal is to understand how task distribution affects the execution time of the Apriori algorithm when applied to different subsets of data.

❖ **Run on one cluster**: When applying the algorithm on a single cluster, we observe an initial execution time, represented by the first bar of the graph. This configuration serves as a reference point for our comparison.

❖ **Run on two clusters:** By increasing the number of clusters to two, the graph illustrates how the distribution of tasks impacts the execution time of the algorithm. We note any significant variation from running on a single cluster.

❖ **Run on three clusters:** Extending our analysis to three clusters, we observe the effect of the increasing complexity of the data distribution on the execution time of the algorithm. This step determines whether task distribution continues to optimize or whether inefficiencies appear with finer distribution.

Table 2 shows the execution time by varying the number of clusters. The graph (as seen in Figure 2) depicts the total execution time of association rule extraction using the Apriori and FP-Growth algorithms as a function of the number of iterations. Apriori is faster than FP-Growth for all iterations.

Table 2: Comparison of execution time for three clusters

Number of clusters	Apriori	FP-Growth
1	0.483	0.855
2	0.538	0.881
3	0.130	0.148

The execution time of FP-Growth is approximately 2 times longer than that of Apriori. The execution time for both algorithms increase with the number of iterations.

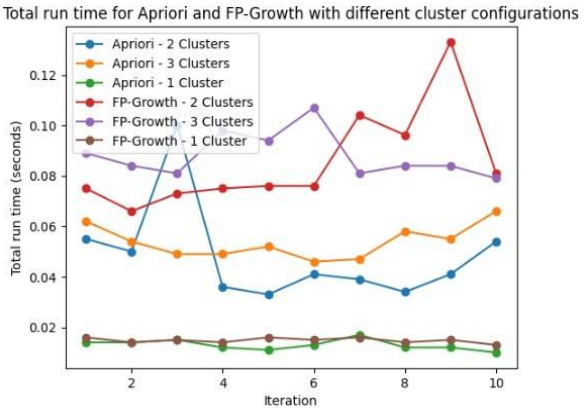

Figure 2: Comparison of apriori algorithm and FP-growth run- time by number of clusters

This is due to the fact that there are more candidates to evaluate at each iteration. The difference in execution time between the two algorithms is more pronounced for later iterations. This is because Apriori is more efficient in handling large candidate sets. Apriori is a more efficient algorithm than FP-Growth for association rule extraction. The performance gap between the two algorithms is more significant for large datasets. The graph (as seen in Figure 3) only shows the total execution time. It does not illustrate the time spent in different phases of the algorithm, such as candidate generation and pruning.

Algorithm performance may vary depending on the size and nature of the dataset. The table 2 shows a comparison of execution times in seconds for the Apriori and FP-Growth algorithms on three different cluster configurations. Cluster counts are listed in the first column, with corresponding runtimes for Apriori and FP-Growth in subsequent columns. It can be observed that for each number of clusters, the execution time of FP-Growth is generally higher than that of Apriori, although the difference is more pronounced with two clusters

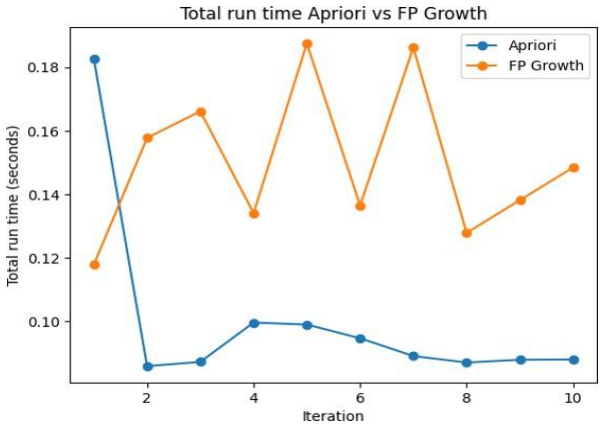

Figure 3: Apriori vs FP-Growth

The graph of association rules between purchases and payments (as seen in Figure 4). The left side of the graph shows the different types of purchases that can be made, while the right side shows the different types of payments that can be made. The lines between the two sides of the graph show the associations between the different purchases and payments. For example, the line between «online purchases" and "credit card" shows that there is a strong association between these two things. This means that people who make online purchases are more likely to pay with a credit card. The numbers on the lines represent the strength of the association. A higher number means that there is a stronger association between the two things. For example, the number on the line between «online purchases" and "Credit card" is 0.0587412587412587, which is a relatively high number. This means that there is a very strong association between online purchases and credit card payments.

5 Conclusion

The preliminary stages of our project were marked by the application of K-Means clustering as an essential phase to group datasets exhibiting similar behaviors. This clustering process created homogeneous clusters, laying the groundwork for a more in-depth exploration of association rules. We adopted a systematic approach by combining K-Means clustering with powerful association rule ex- traction algorithms, namely Apriori and FP-Growth. These were chosen for their distinct capabilities: Apriori's strategy of generating and pruning candidate sets, and FP-Growth's efficient tree construction.

The choice between Apriori and FP-Growth was guided by considerations specific to our dataset, highlighting the importance of understanding the unique characteristics of each algorithm and adapting accordingly. While Apriori follows an incremental approach to generating and pruning, FP-Growth employs recursive exploration with FP tree construction. Our integrated approach facilitated a thorough exploration of specific behaviors within each cluster, thanks to the application of Apriori on datasets al- ready grouped by K-

Means. This synergy between clustering and association rule extraction strengthened the relevance of the obtained results. The project focused on refining association rule extraction techniques; especially in the context of exploring non-distributed datasets. Although we did not introduce entirely new methods, our improvements make a significant contribution to addressing challenges posed by large-scale datasets. Our study underscores the crucial importance of effectively leveraging established algorithms, paving the way for more efficient data exploration practices.

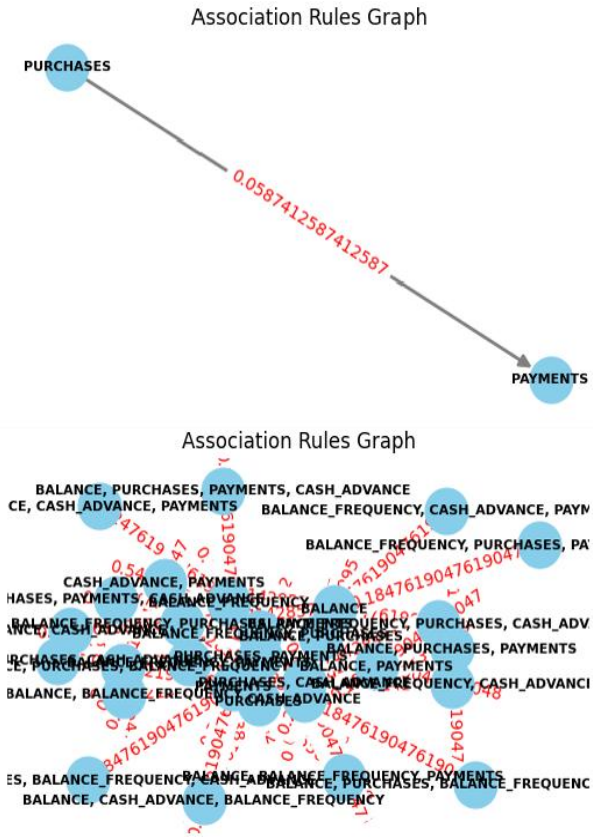

Figure 4: The graph of association rules

References

[1] Hassan Ibrahim Hayatu, Abdullahi Mohammed, and Ahmad Barroon Ismaeel. Big data clustering techniques: Recent advances and survey. Machine Learning and Data Mining for Emerging Trend in Cyber Dynamics, pages 57–79, 2021. doi:10.1007/978-3-030-66288-2_3.

[2] A. K. Jain, A. Topchy, M. H. C. Law, and J. M. Buhmann, “Landscape of clustering algorithms,” in Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004., Jan. 2004.DOI: 10.1109/icpr.2004.1334073

[3] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann, “Stability-Based Validation of Clustering Solutions,” Neural Computation, vol. 16, no. 6, pp. 1299–1323, Jun. 2004. DOI: 10.1162/089976604773717621

[4] F. Li, J. Qin, and Y. Kang, “Closed-Loop Hierarchical Operation for Optimal Unit Commitment and Dispatch in Microgrids: A Hybrid System Approach,” arXiv.org, 2018. https://arxiv.org/abs/1812.09928

[5] H. Yin, A. Aryani, S. Petrie, A. Nambissan, A. Astudillo, and S. Cao, “A Rapid Review of Clustering Algorithms,” arXiv.org, Jan. 2024, https://arxiv.org/abs/2401.07389

[6] M. H. Hansen and B. Yu, “Model Selection and the Principle of Minimum Description Length,” Journal of the American Statistical Association, vol. 96, no. 454, pp. 746–774, Jun. 2001. DOI: 10.1198/016214501753168398

[7] Youmeng Guan, An Algorithm for Data Management of Higher Education Based on Fuzzy Set Theory - Association Rule Mining Algorithm, Informatica 45 (2021) 157–164, https://doi.org/ 10.31449/inf. v47i9. 5222

[8] Cheung, David W., et al. "A fast distributed algorithm for mining association rules." Fourth international conference on parallel and distributed information systems. IEEE, 1996.

[9] Fernandez-Basso, Carlos, M. Dolores Ruiz, and Maria J. Martin-Bautista. "New spark solutions for distributed frequent itemset and association rule mining algorithms." Cluster Computing 27.2 (2024): 1217-1234.

[10] Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." IEEE Transactions on neural networks 16.3 (2005): 645-678.

[11] Mining, What Is Data. Introduction to data mining. New Jersey: Pearson Education, Inc, 2006.

[12] Zaharia, Matei, et al. "Spark: Cluster computing with working sets." 2nd USENIX workshop on hot topics in cloud computing (HotCloud 10). 2010.

[13] Shaik, Indivar, et al. "K-means clustering algorithm based on e-commerce big data." International Journal of Innovative Technology and Exploring Engineering 8.11 (2019): 1910-1914.

[14] Putri, Indah Dwi, and Romi Hardianto. "Application of the FP-Growth Algorithm in Consumer Purchasing Pattern Analysis." Journal of Computer Scine and Information Technology (2024): 44-49.

[15] S. Guha, R. Rastogi, and K. Shim, “Rock: A robust clustering algorithm for categorical attributes,” Information Systems, vol. 25, no. 5, pp. 345–366, Jul. 2000. DOI: 10.1016/s0306-4379(00)00022-3

[16] Aida Chefrour and Samia Drissi, K-CAE: Image Classification Using Convolutional AutoEncoder Pre-Training and K-means Clustering, Informatica 47 (2023), https://doi.org/10.31449/inf.v46i7.4499

[17] K. Sharma, S. Saini, S. Sharma, H. S. Kang, M. Bouye, and D. Krah, "Big Data Analytics Model for Distributed Document Using Hybrid Optimization with K-Means Clustering," Wireless Communications and Mobile Computing, vol. 2022, p. e5807690, Jun. 2022. DOI: 10.1155/2022/5807690.

[18] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 09.011 2010. DOI: 10.1016/j.patrec.2009.

[19] C. C. Aggarwal, C. K. Reddy, and T. Francis, Data Clustering Algorithms and Applications. Boca Raton, FL: Chapman And Hall/CRC, 2018. https://github. com/RAHULKASHYAP02/CreditCardSegmentation / blob/master/CC%20GENERAL.csv