Extraction of association rules in a diabetic dataset using parallel FP-growth algorithm under apache spark

Youssef Fakir, Salim Khalil, Mohamed Fakir
Department of Computer Sciences, Faculty of Sciences and Technics, Sultan Moulay Slimane University, Beni Mellal, Morocco

Article Info

Article history:

Received Jan 6, 2024
Revised May 12, 2024
Accepted Jun 17, 2024

Keywords:

Apache spark
Association rules
Diabetes prediction
FP-growth
Parallel FP-growth

ABSTRACT

This research paper focuses on enhancing the frequent pattern growth (FP-growth) algorithm, an advanced version of the Apriori algorithm, by employing a parallelization approach using the Apache Spark framework. Association rule mining, particularly in healthcare data for predicting and diagnosing diabetes, necessitates the handling of large datasets which traditional methods may not process efficiently. Our method improves the FP-growth algorithm's scalability and processing efficiency by leveraging the distributed computing capabilities of apache spark. We conducted a comprehensive analysis of diabetes data, focusing on extracting frequent itemsets and association rules to predict diabetes onset. The results demonstrate that our parallelized FP-growth (PFP-growth) algorithm significantly enhances prediction accuracy and processing speed, offering substantial improvements over traditional methods. These findings provide valuable insights into disease progression and management, suggesting a scalable solution for large-scale data environments in healthcare analytics.

Corresponding Author:

Youssef Fakir
Department of Computer Sciences, Faculty of Sciences and Technics, Sultan Moulay Slimane University
Beni Mellal, Morocco
Email: info.dec07@yahoo.fr

1. INTRODUCTION

In the realm of data mining, the extraction of meaningful patterns from large and complex datasets is crucial, particularly in healthcare, where precise analysis can significantly influence outcomes. This discipline plays a pivotal role in transforming vast amounts of raw data into useful information, which is particularly essential in fields requiring detailed data analysis [1]. Among various data mining techniques, this paper focuses on association analysis [2], [3], classification [4], and clustering [5], [6], which are increasingly vital due to the burgeoning volume of data generated in medical research and practice. Traditional algorithms often falter under the weight of such datasets, as they require substantial computational power and storage capacity, leading to inefficiencies in mining frequent patterns [7].

Association rule mining, a key method within data mining, involves identifying frequent itemsets and using these to generate strong association rules [8]. This process typically consists of two major steps: the first is the identification of all frequent itemsets where the support is at least as great as the specified minimum support threshold, and the second is the generation of robust association rules from these itemsets that satisfy predefined minimum support and confidence levels [9]. Despite the importance of association rule mining, the efficiency of traditional methods like the Apriori algorithm degrades significantly with large data volumes [10]. These methods are constrained by their computational intensity and are often incapable of handling large-scale data effectively, as evidenced by their slow processing times and high memory usage.

Addressing these challenges, this paper introduces a novel approach by parallelizing the frequent pattern growth (FP-growth) algorithm using the apache spark framework. This method, designed to overcome the limitations of traditional algorithms, leverages horizontal partitioning to enhance adaptability and processing efficiency. Our implementation of parallelized FP-growth (PFP-growth) algorithm significantly improves computation time, memory usage, and the efficiency of identifying frequent itemsets, which is critical for effective data analysis in healthcare settings [11].

Our implementation of parallelized FP-growth (PFP-growth) algorithm significantly improves computation time, memory usage, and the efficiency of identifying frequent itemsets, which is critical for effective data analysis in healthcare settings [12]. This study aims to harness the advanced capabilities of the PFP-growth algorithm to extract and analyze association rules from a comprehensive diabetes dataset, thereby enhancing the predictive accuracy and understanding of disease progression [13]. By integrating the PFP-growth algorithm into the spark platform, we propose a scalable, efficient solution to analyze large-scale healthcare data, offering new insights into diabetes management and risk factor identification [14]. The following sections will review relevant literature to further contextualize our approach, describe our methodology, present the dataset transformation and results, and discuss the implications of our findings in improving diabetes prediction and management.

2. LITTERATURE REVIEW

Over the last few decades, the development of algorithms for extracting association rules from datasets has significantly advanced, reflecting a core area of data mining research [15]. The drive to accelerate the discovery of association rules has often been motivated by the need to improve the efficiency of mining operations, particularly as datasets grow and complexity.

Initially, he focuses was on foundational algorithms such as the apriori algorithm, which remains a cornerstone in the field for its basic yet effective approach to rule discovery. The Apriori algorithm and its derivatives excel in efficiently generating frequent itemsets, a fundamental and computationally intensive step in association rule mining [16].

Recent advancements have included specialized algorithms that enhance efficiency and applicability. For instance, an efficient algorithm tailored for Spark has shown promise in handling large-scale data more effectively [17]. The R-apriori algorithm is another adaptation that optimizes the mining of frequent items under specific constraints [18]. The utility of association rules extends beyond traditional applications; for example, a method utilizing moodle activity log data for cluster and association analysis visualization has been developed, demonstrating the versatility of association rules in various contexts [19].

In practical applications, association rules are pivotal in diverse fields such as healthcare and genetics. They are employed for tasks ranging from the detection and monitoring of infectious diseases [20], [21], understanding medication prescription patterns [22], to uncovering genetic patterns [23]. Specifically, in healthcare, association rules have been used to identify risk factors in diabetes, a critical area of study given the global prevalence of the condition [24]. Furthermore, these techniques have been applied to pediatric data to detect common risk factors in diseases affecting children [25]. The application of fuzzy sets to extend the functionality of association rules illustrates the ongoing innovation in the field. This approach allows for handling imprecise data, thereby broadening the scope of environments in which association rule mining can be effectively applied.

Despite these advancements, traditional algorithms for extracting association rules often struggle with time and memory efficiency. This has led to the exploration of parallel algorithms as a necessary evolution to address the computational demands of large datasets. The proposed study utilizes the PFP algorithm on a diabetes dataset to extract itemsets that are indicative of risk factors leading to the disease, showcasing a practical application of these theoretical advancements. This literature review sets the stage for discussing the implementation of the PFP algorithm in the subsequent section, highlighting its potential to address the identified challenges and improve the efficiency of association rule mining in large-scale healthcare data environments.

3. PROPOSED METHOD

This study introduces the PFP-growth algorithm, a robust solution for extracting association rules from large-scale datasets. Leveraging the Apache Spark framework, this innovative approach addresses the scalability challenges of the traditional FP-growth algorithm. By overcoming these inherent limitations, the PFP-growth algorithm is particularly effective in handling expansive data.

3.1. Horizontal partitioning and algorithm adaptation

The conventional FP-growth algorithm, renowned for its efficiency in smaller datasets, utilizes a compact tree structure to capture transaction patterns without candidate generation. While offering significant advantages over the apriori algorithm, FP-growth struggles with scalability, often facing significant slowdowns and increased memory usage as data volume grows. To mitigate these scalability issues, we employ horizontal partitioning, a technique that divides the data across multiple computational nodes within the spark cluster. This segmentation facilitates independent processing on manageable data chunks, enhancing both scalability and processing speed.

We have adapted the FP-growth algorithm to effectively harness spark’s distributed computing model. Each node constructs a local FP-tree from its subset of the dataset and independently executes the mining process. This parallel structure allows for localized frequent itemset generation without the overhead of centralized data processing.

These modifications ensure that the PFP-growth algorithm not only maintains the integrity of the mining process but also achieves substantial gains in efficiency. By leveraging multiple nodes, the workload is distributed, significantly reducing the time required for mining operations. The parallel nature of this approach also limits the memory load on any single node, thereby preventing bottlenecks and ensuring faster computation.

3.1.1. Detailed execution within spark’s architecture

Apache spark’s architecture is integral to the implementation of the PFP-growth algorithm. Operating under a master-slave setup, spark employs a central coordinator (master) to manage the distribution of tasks and resources across multiple worker nodes (slaves). Here are the execution process steps as demonstrated in Figure 1.

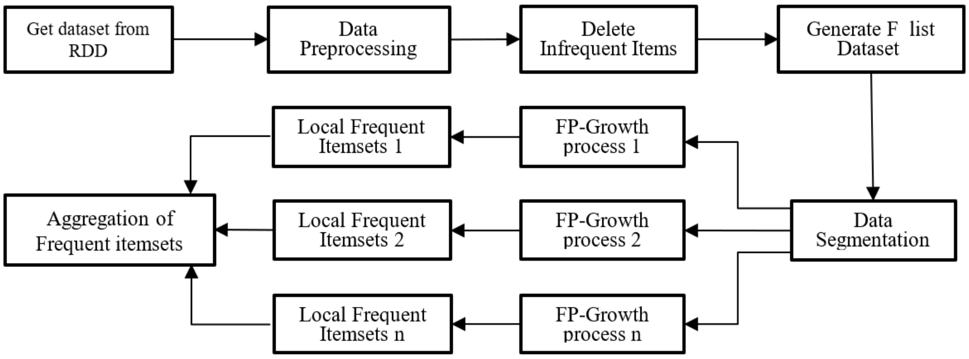


Figure 1. PFP-growth algorithm execution process

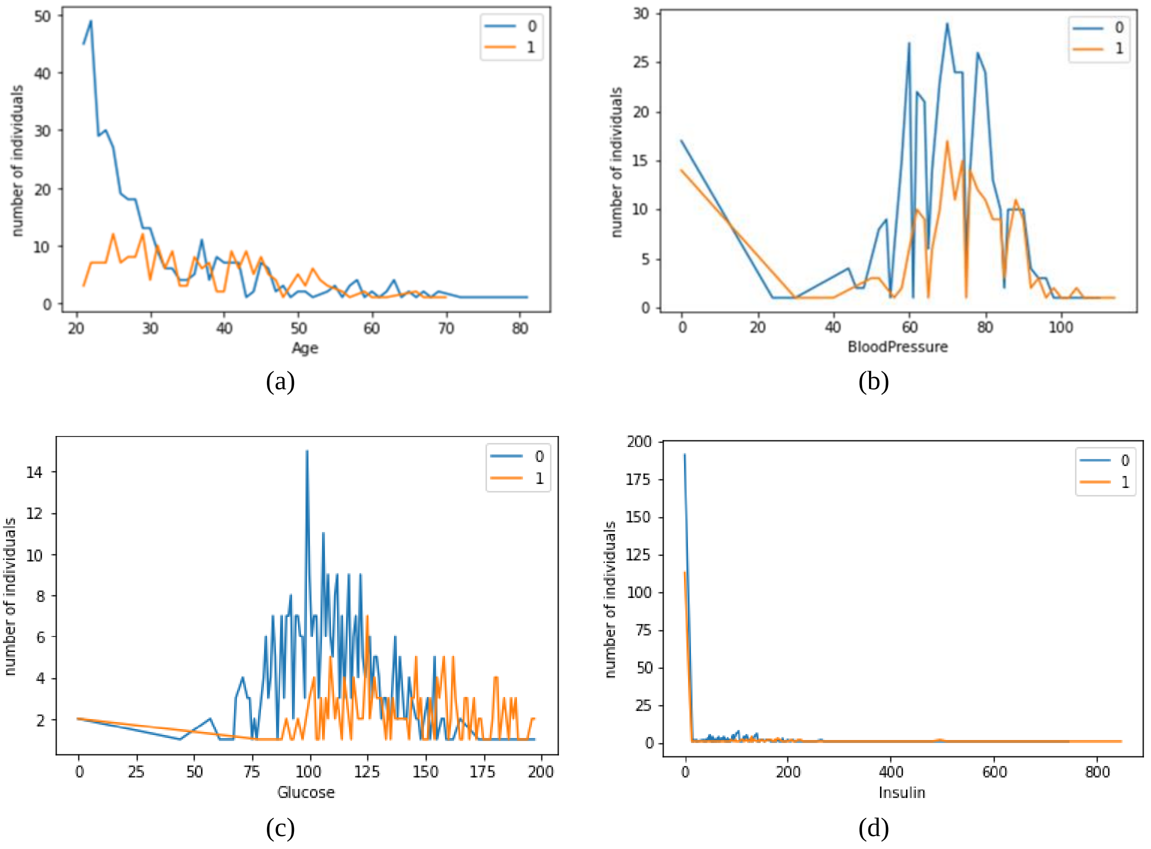
- Step 1: data distribution- initially, the complete dataset is distributed as a resilient distributed dataset (RDD). This step sets the foundation for high availability and robust data handling.
- Step 2: data preprocessing- data is pre-processed and infrequent items are removed. This step enhances the efficiency of the subsequent mining process by reducing the dataset size and complexity.
- Step 3: data segmentation- each worker node receives segments of the F_list, stored as RDDs, ready for local processing. This segmentation ensures that data handling is manageable and scalable across nodes.
- Step 4: local mining operations- workers independently execute the FP-growth process to identify frequent itemsets based on predefined support thresholds.
- Step 5: intermediate aggregation- post mining, each node aggregates its frequent itemsets, labeled as local frequent itemsets. This aggregation prepares the data for a unified analysis across the spark cluster.
- Step 6: global synthesis– results are aggregated and synthesized from all nodes to form the comprehensive set of association rules. This ensures that the outcomes are accurate and cover all possible frequent patterns within the dataset.

The utilization of RDDs not only facilitates efficient data manipulation and fault recovery but also supports the parallel execution of tasks. Spark’s execution model, encompassing task distribution and resource management, is ideally suited to meet the demands of large-scale data mining. This architecture provides a robust framework for the PFP-growth algorithm, ensuring that it performs optimally across varied and large datasets. Once the architecture is set up, the next crucial step involves preparing the actual data for mining, as detailed in the following section.

3.2. Data preparation and transformation for mining

To effectively utilize the PFP-growth algorithm within the apache spark framework, our study required meticulous preparation and transformation of the Pima Indians Diabetes database [26]. This well-curated dataset, which includes records from 768 women of Pima Indian descent, contains several critical attributes such as plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin levels, body mass index (BMI), diabetes pedigree function, the number of pregnancies, and age. These attributes are systematically documented in Table 1, providing a structured overview that highlights each variable's range. Nevertheless, it is essential to transform these continuous variables into categorical intervals, which are crucial for identifying and leveraging patterns predictive of diabetes in our algorithm.

Guided by rigorous statistical analysis and substantial domain knowledge, we transformed these attributes into categorical intervals. This strategic categorization is designed to distinguish between diabetic and non-diabetic groups based on specific attributes. For instance, age was categorized into 'Young' [0, 30] and 'senior' [31, 80] groups to reflect different risk profiles. This categorization is visualized in Figure 2, Figure 2(a), which illustrates the age distribution of diabetic and non-diabetic individuals. Diabetic patients are denoted with an orange color and the symbol "0," while non-diabetic individuals are marked in blue and represented by the symbol "1." Blood pressure readings were segmented into low [0, 40], medium [41, 90], and high [91, 120] categories to correlate with varying diabetes risks. This distribution is shown in Figure 2(b). Similarly, glucose levels were divided into normal [0, 125] and high [126, 200], aligning with clinical thresholds for diabetes diagnosis, as depicted in Figure 2(c). Insulin levels were also split into low [0, 30], medium [31, 150], and high [151, 800] categories to capture the variations in insulin dynamics, detailed in Figure 2(d). These transformations, as detailed in Table 1, were applied to all remaining attributes for the subsequent data analysis phase. This enabled precise handling and effective mining of the dataset to accurately predict the onset of diabetes. This strategic categorization, detailed in Table 1 and illustrated in figures showing the distribution of these attributes among diabetic and non-diabetic individuals, ensures the dataset is optimally structured for our FP-growth implementation on apache spark's distributed computing platform.



(a)

(b)

(c)

(d)

Figure 2. Age distribution; (a) blood pressure distribution, (b) glucose levels distribution, (c) insulin levels distribution, and (d) of diabetic and non-diabetic individuals

Table 1. Descriptive summary and categorical transformations for diabetes prediction

Attributes	Description	Value interval	Categorical intervals
P	It shows how many times patient is pregnant	[0-17]	P1 {0-5}, P2{>5}
G	Plasma glucose concentration over 2 h in an oral glucose tolerance test	[0-199]	G1{0-125}, G2{>125}
BP	It indicates the patient's blood pressure (mm Hg)	[0-122]	B1{0-40}, B2{40-90}, B3{>90}
S	It shows skin fold thickness	[0-99]	S1{0-8}, S2{8-45}, S3{>45}
I	2-Hour serum insulin (mu U/ml)	[0-846]	I1{0-30}, I2{30-150}, I3{>150}
BMI	It indicates body mass index	[0-67]	BMI1{0-30}, BMI2{>30}
D	It shows family history of patient	[0-2.45]	D1{0-0.8}, D2{>0.8}
A	It shows age of patient	[21-81]	A1 {0-30}, A2{>30}
O	1 for diabetes and 0 for non-diabetes	(0,1)	0 for non-diabetic, and 1 for diabetic

4. RESULTS AND DISCUSSION

4.1. Analysis of PFP-growth performance

The comparative analysis of the computing times for the PFP-growth, FP-growth, and R-apriori algorithms, as depicted in Figure 3, reveals significant performance differences across three runs. The PFP-growth algorithm demonstrates a consistent reduction in computing time from 6 seconds in the first run to 4 seconds in the third, highlighting its efficient scalability and optimization when parallelized using the apache spark framework. In contrast, the FP-growth algorithm shows relatively higher computing times, decreasing marginally from 18.38 seconds in the first run to 17.48 seconds in the third. This variation suggests that while FP-growth benefits from parallel processing, it does not achieve the same efficiency gains as the PFP-growth algorithm.

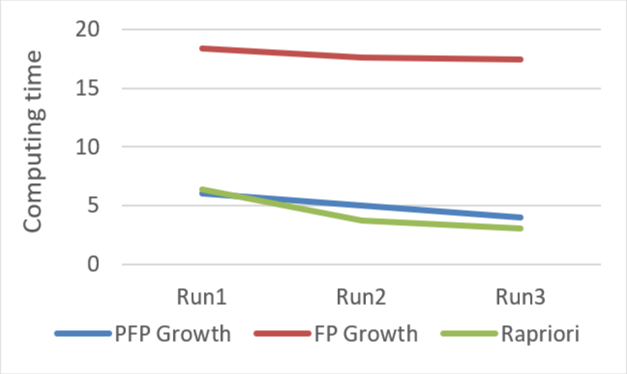


Figure 3. Comparative computing time of PFP-growth, FP-growth, and R-apriori algorithms across three runs

Interestingly, the R-apriori algorithm starts with a computing time of 6.35 seconds, which significantly reduces to 3.09 seconds by the third run. This performance indicates that R-apriori optimizes its process more effectively over successive runs, likely due to better handling of datasets with constraints or more efficient data pruning techniques. These findings underscore the enhanced capability of the PFP-growth algorithm to manage large datasets efficiently, a crucial advantage for its application in healthcare data analytics for diabetes prediction. The parallelization of the FP-growth algorithm appears to effectively reduce computational overhead, thereby improving processing speed considerably compared to its non-parallel counterpart.

4.2. Performance of the parallel FP-growth algorithm in extracting itemsets

Figure 4 illustrates how the PFP-growth algorithm's performance varies with different minimum support thresholds. There is a notable decrease in the number of itemsets generated as the minimum support threshold increases. At a lower threshold (e.g., 0.2), the algorithm identifies a larger number of itemsets, capturing more granular patterns within the dataset. However, this number significantly diminishes as the threshold increases, indicating that fewer itemsets meet the higher criteria of support.

This trend highlights the algorithm's ability to filter out less significant itemsets and focus computational resources on those patterns that occur more frequently. Such behavior is vital for efficiently managing large-scale data environments like those in healthcare analytics, where identifying the most impactful patterns can substantially affect predictive accuracy and patient outcomes. The relationship between the number of itemsets and the minimum support threshold also reflects the algorithm’s adaptability to various data densities and distributions, which is essential for tailoring the pattern discovery process to specific research needs or clinical relevance.

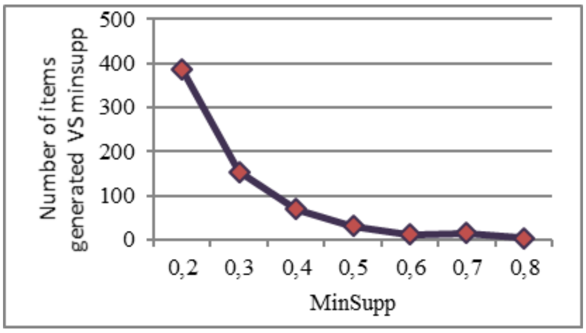


Figure 4. Performance of parallel FP-growth algorithm in extracting itemsets at different minimum support thresholds

4.3. Analysis of association rules for diabetes prediction

The top 10 association rules extracted from the diabetes dataset, generated with a minimum support of 0.4 and a confidence level of 0.5, offer vital insights into factors significantly associated with diabetes risk. These rules, detailed in Table 2, integrate various patient characteristics such as age, body mass index (BMI), glucose levels, family history, blood pressure, skin fold thickness, and insulin levels. They reveal patterns that either increase or decrease the likelihood of developing diabetes. For instance, rule 1 indicates that younger patients (A1: age≤30) with a BMI≤30 (BMI1), fewer pregnancies (P1:≤5), and a lower family diabetes history (D1:≤0.8) are less likely to have diabetes. This suggests that traditional risk factors such as age, obesity, and family history significantly influence the onset of diabetes. Additionally, rule 2 highlights that patients with high glucose levels (G2:>125 mg/dl) and a higher BMI (BMI2:>30) are at increased risk for diabetes, aligning with established medical knowledge that links elevated glucose levels and obesity with heightened diabetes risk.

Table 2. Key association rules identifying risk and protective factors for diabetes

Rule number	Antecedent	Consequent	Confidence
1	[“BMI1”, “A1”, “P1”, “D1”]	[“0”]	0.93
2	[“G2”, “BMI2”]	[“1”	0.90
3	[“G1”, “S2”, “P1”, “D1”, “B2”]	[“0”]	0.89
4	[“A2”, “G2”, “BMI2”]	[“1”]	0.88
5	[“G1”, “S2”, “D1”]	[“0”]	0.85
6	[“A1”, “P1”, “G1”, “S1”]	[“0”]	0.85
7	[“S3”, “G2”]	[“1”]	0.85
8	[“I3”, “G2”]	[“1”]	0.83
9	[“B3”, “BMI2”]	[“1”]	0.80
10	[“G2”, “BMI2”]	[“B2”]	0.859813084

In conclusion, the key insights derived from these rules include:

- Protective factors: younger age (≤ 30 years), lower BMI (≤30), fewer pregnancies (≤5), and lower family history scores (≤0.8) collectively contribute to a lower likelihood of diabetes. These factors suggest that proactive management of lifestyle and genetic predispositions from a young age can significantly reduce diabetes risk.
- Intermediate risk factors: moderate skin fold thickness (8-45 mm) and blood pressure (40-90 mm Hg), when appearing alongside other moderate risk factors, still indicate a generally lower risk for diabetes, suggesting that these factors alone do not significantly raise diabetes risk unless accompanied by more severe indicators.
- High-risk profiles: high glucose levels (>125 mg/dl) coupled with higher BMI (>30) emerge as strong predictors of diabetes. This confirms well-established clinical understandings that link metabolic dysfunction with the development of diabetes.

- Compounded risk from hypertension and obesity: individuals with very high blood pressure (>90 mm Hg) and higher BMI are significantly more likely to develop diabetes, underscoring the interplay between hypertension, obesity, and metabolic health.

These findings notably contribute to medical knowledge by enhancing predictive models and informing targeted interventions. By integrating these insights, healthcare providers can develop personalized management plans that address specific risk factors identified in patients, thus improving preventive and therapeutic approaches in diabetes care. This analysis not only deepens our understanding of the multifactorial nature of diabetes but also underscores the value of advanced data mining techniques in discovering complex patterns within medical datasets.

5. CONCLUSION

This study underscores the substantial benefits of utilizing the PFP-growth algorithm within apache spark for mining association rules from large healthcare datasets. By exploiting spark’s distributed computing capabilities, we have significantly improved the scalability and efficiency of the FP-growth algorithm, facilitating faster and more precise analysis of complex datasets like those encountered in diabetes research. The PFP-growth algorithm’s adept handling of large data volumes with greater speed and reduced memory overhead represents a marked improvement over traditional method like the apriori algorithm. The insights derived from the diabetes dataset not only deepen our understanding of the disease’s dynamics but also demonstrate the potential applicability of this method to other medical research areas with similar data challenges. Moreover, the analysis of association rules has led to the identification of critical risk and protective factors for diabetes, providing a foundation for more targeted medical interventions and personalized treatment approaches. In essence, integrating advanced data mining techniques with robust platforms like apache spark is transforming the landscape of healthcare research, offering new avenues for enhancing disease management and medical diagnostics.

REFERENCES

[1] J. P. J. Han, M. Kamber, “Data mining concepts and techniques third edition,” *Paper Knowledge . Toward a Media History of Documents*, pp. 12–26, 2020.

[2] C. Gyorodi, R. Gyorodi, T. Cofeey, and S. Holban, “Mining association rules using dynamic FP-trees,” *Proceedings of irish signals and systems conference*, pp. 76–81, 2003.

[3] G. Agapito, P. H. Guzzi, and M. Cannataro, “Parallel extraction of association rules from genomics data,” *Applied Mathematics and Computation*, vol. 350, pp. 434–446, Jun. 2019, doi: 10.1016/j.amc.2017.09.026.

[4] M. Antonie, A. Coman, and O. R. Zaiane, “Application of data mining techniques for medical image classification,” *Proceedings of the second international Workshop on Multimida Data Mining (MDM/KDD’2001)*, pp. 94–101, 2001, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.9742&rep=rep1&type=pdf.

[5] B. Lorbeer, A. Kosareva, B. Deva, D. Softić, P. Ruppel, and A. Küpper, “Variations on the clustering algorithm BIRCH,” *Big Data Research*, vol. 11, pp. 44–53, Mar. 2018, doi: 10.1016/j.bdr.2017.09.002.

[6] Y. Fakir and J. Elklil, “Clustering techniques for big data, lecture notes in business information processing,” *springer*, 2021.

[7] Huan Liu and Lei Yu, “Toward integrating feature selection algorithms for classification and clustering,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, Apr. 2005, doi: 10.1109/TKDE.2005.66.

[8] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," ACM SIGMOD Record, vol. 22, no. 2, pp. 207-216, 1993.

[9] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," ACM SIGMOD Record, vol. 29, no. 2, pp. 1-12, 2000.

[10] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, 2013.

[11] T. T. Nguyen, “A compact FP-tree for fast frequent pattern retrieval,” *27th Pacific Asia Conference on Language, Information, and Computation, PACLIC 27*, pp. 430–439, 2013.

[12] J. Xu, Y. Wang, and Z. Zhang, "A Novel Parallel Algorithm for Frequent Itemset Mining in Big Data," IEEE Access, vol. 8, pp. 135973-135982, 2020.

[13] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019.

[14] Y. Fakir, R. Elayachi, and B. Mahi, “Clustering objects for spatial data mining: a comparative study,” *Journal of Big Data Research*, vol. 1, no. 3, pp. 1–11, Mar. 2023, doi: 10.14302/issn.2768-0207.jbr-23-4478.

[15] H. Qiu, R. Gu, C. Yuan, and Y. Huang, “YAFIM: a parallel frequent itemset mining algorithm with spark,” in *Proceedings - IEEE 28th International Parallel and Distributed Processing Symposium Workshops, IPDPSW 2014*, May 2014, pp. 1664–1671, doi: 10.1109/IPDPSW.2014.185.

[16] B. Wu, D. Zhang, Q. Lan, and J. Zheng, “An efficient frequent patterns mining algorithm based on apriori algorithm and the FP-tree structure,” in *Proceedings - 3rd International Conference on Convergence and Hybrid Information Technology, ICCIT 2008*, Nov. 2008, vol. 1, pp. 1099–1102, doi: 10.1109/ICCIT.2008.109.

[17] Y. Xun, J. Zhang, H. Yang, and X. Qin, “HBPFP-DC: a parallel frequent itemset mining using spark,” *Parallel Computing*, vol. 101, p. 102738, Apr. 2021, doi: 10.1016/j.parco.2020.102738.

[18] S. Uthra and K. Rohini, “An efficient R-apriori algorithm for frequent item set mining in python,” *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 3516–3519, Jul. 2019, doi: 10.35940/ijrte.B3024.078219.

[19] A. R. Tamba, K. Lumbantoruan, A. Pakpahan, and S. Situmeang, “A cluster and association analysis visualization using moodle activity log data,” *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 12, no. 2, p. 150, Aug. 2023, doi: 10.11591/ijict.v12i2.pp150-161.

[20] C. Creighton and S. Hanash, “Mining gene expression databases for association rules,” *Bioinformatics*, vol. 19, no. 1, pp. 79–86, Jan. 2003, doi: 10.1093/bioinformatics/19.1.79.

[21] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser, “Association rules and data mining in hospital infection control and public health surveillance,” *Journal of the American Medical Informatics Association*, vol. 5, no. 4, pp. 373–381, Jul. 1998, doi: 10.1136/jamia.1998.0050373.

[22] T.-J. Chen, L.-F. Chou, and S.-J. Hwang, “Application of a data-mining technique to analyze coprescription patterns for antacids in Taiwan,” *Clinical Therapeutics*, vol. 25, no. 9, pp. 2453–2463, Sep. 2003, doi: 10.1016/S0149-2918(03)80287-4.

[23] M. Delgado, D. Sánchez, M. J. Martín-Bautista, and M.-A. Vila, “Mining association rules with improved semantics in medical databases,” *Artificial Intelligence in Medicine*, vol. 21, no. 1–3, pp. 241–245, Jan. 2001, doi: 10.1016/S0933-3657(00)00092-0.

[24] P. Tiwari and V. Singh, “Diabetes disease prediction using significant attribute selection and classification approach,” *Journal of Physics: Conference Series*, vol. 1714, no. 1, p. 012013, Jan. 2021, doi: 10.1088/1742-6596/1714/1/012013.

[25] S. M. Downs and M. Y. Wallace, “Mining association rules from a pediatric primary care decision support system.,” *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 200–204, 2000.

[26] “PIMA Indian dataset,” *kaggle*, 2021. https://www.kaggle.com/uciml/pima-indians-diabetes-database.

BIOGRAPHIES OF AUTHORS

Youssef Fakir received his master’s degree in business intelligence and Ph.D. degree in Computer Sciences from Sultan Moulay Slimane University, Beni Mellal, Morocco, respectively in 2019 and 2023. He is currently a data analyst within CNSS. His main research interests focus on artificial intelligence, information retrieval, feature selection, data mining, and text mining. He can be contacted at email: info.dec07@yahoo.fr.

Salim Khalil is a computer scientist with expertise in web mining. He received his Ph.D. degree from the Faculty of Sciences and Techniques in Beni Mellal in 2021, Morocco, where he developed Rcrawler, a web scraping tool using the R language. His research has been published in respected scientific journals, highlighting his contributions to data analytics and web mining. His research interests include machine learning, datamining, and artificial intelligence. He can be contacted at email: khalilsalim1@gmail.com.

Mohamed Fakir received his master’s in Electronic Electric System Engineering from Nagaoka University of Technology in 1991 and Ph.D. degree in Computer Sciences from Cadi Ayyad University in 2001. He was a staff with Hitachi Ltd., Japan from 1991 to 1994 at air conditioning department. He is currently with the Department of Computer Sciences of Faculty of Science and Technics de Beni Mellal, Morocco. He was the general chair of five edition of the international conference on business intelligence. His research interests include machine learning, datamining, and artificial intelligence. He can be contacted at email: fakfad@yahoo.fr.