International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET

Efficiency and Effectiveness in Utility-Based and Frequent Itemset Mining: A Comprehensive Review

Nishigandha Mhatre[1]
Department of Computer Engineering, Pillai HOC College of Engineering and Technology, Rasayani,
Maharashtra, India
E-mail: nishi.nmhatre@gmail.com

Srijita Bhattacharjee[2]
Department of Computer Engineering, Pillai HOC College of Engineering and Technology, Rasayani,
Maharashtra, India
E-mail: srijitab@mes.ac.in

ARTICLEINFO

Article History :
Accepted: 02 Feb 2024
Published: 07 Feb 2024

Publication Issue :
Volume 11, Issue 1
January-February-2024
Page Number :
179-192

ABSTRACT

The rapid growth of data in various domains has led to the need for efficient pattern mining algorithms that can handle large-scale datasets. In this study, the comprehensive review in the domains of utility-based and frequent itemset mining (FIM), focusing on the dual facets of efficiency and effectiveness. In the ever-expanding landscape of data mining and knowledge discovery, the extraction of valuable patterns and insights from large datasets is paramount. Utility-based mining addresses the challenge of incorporating user-defined measures of importance, reflecting real-world applications where not all items are equal. Simultaneously, FIM seeks to identify recurring patterns within datasets, providing valuable associations and dependencies. This review synthesizes recent advancements and methodologies in both utility-based and FIM, analyzing their respective strengths and limitations. Efficiency considerations encompass algorithmic optimizations, parallel computing, and scalability, ensuring that the mining process is computationally tractable for large-scale datasets. Effectiveness evaluations delve into the quality of discovered patterns, emphasizing their relevance and utility in diverse applications. The synthesis of these two mining paradigms underscores the importance of striking a balance between computational efficiency and the ability to extract meaningful patterns. By examining state-of-the-art techniques and methodologies, this review aims to provide researchers, practitioners, and decision-makers with a comprehensive understanding of the current landscape in utility-based and FIM, offering insights into future directions for advancing the efficiency and effectiveness of pattern discovery in diverse domains.

Keywords - Frequent Mining, Utility Mining, Performance Parameters, Itemset Mining.

I. INTRODUCTION

Data analysis has emerged as a crucial field in data mining, with applications spanning various industries, including aviation, medicine, and manufacturing. In the area of data mining, the extraction of meaningful patterns and associations from large datasets is a critical endeavor. Two primary paradigms that have emerged to address this challenge are utility-based mining and FIM. Utility-based mining extends traditional association rule mining by considering the importance or utility of items in transactions, mirroring real-world scenarios where items exhibit varying degrees of significance. Concurrently, FIM focuses on the identification of sets of items that frequently co-occur in datasets, laying the groundwork for the derivation of association rules. The surge in data complexity and volume necessitates a nuanced understanding of the efficiency and effectiveness of mining algorithms within these paradigms. Efficiency denotes the computational prowess of algorithms in handling vast datasets, while effectiveness gauges the quality and relevance of the patterns discovered. Given the multitude of algorithms available for both utility-based and FIM, a comprehensive review becomes imperative to distill insights into their comparative performance across various parameters.

This review provides a panoramic view of the landscape, scrutinizing the intricacies of utility-based and FIM algorithms. Notably, our focus lies in systematically comparing their efficiency and effectiveness on multiple performance parameters. These parameters encompass computational complexities, scalability, handling of diverse data types, and the ability to adapt to evolving patterns. By delving into the strengths and limitations of each algorithm, we aim to guide researchers, practitioners, and decision-makers in selecting the most suitable approach based on the specific demands of their applications. Through this comprehensive exploration, we aspire to contribute to the ongoing discourse in the field, fostering a deeper understanding of the intricate trade-offs between efficiency and effectiveness in utility-based and FIM. As we navigate through the intricacies of these paradigms, the comparative analysis of algorithms will shed light on the state-of-the-art methodologies and pave the way for future advancements in the pursuit of extracting valuable insights from complex and voluminous datasets.

II. LITERATURE REVIEW

Mining high-utility association rules from transactional databases has gained significant attention due to its potential for discovering valuable patterns and insights in various domains. High-utility itemset mining is a specialized form of itemset mining that focuses on identifying itemsets with high utility values, often associated with transactions or items' profits, rather than just frequent itemsets. This task has applications in recommendation systems, market basket analysis, and resource allocation. A literature review on HUIM algorithms is presented, encompassing researcher’s notable works:

Yang et al. [1] introduced a HUIM algorithm based on Particle Filter, leveraging a stochastic sampling method to efficiently explore the search space. Their approach utilizes a particle filter to estimate the utility of itemsets, enabling the identification of HUI without requiring a predefined threshold. By dynamically adjusting the particle filter parameters, the algorithm efficiently identifies HUI, making it suitable for large-scale datasets.

Liu et al. [2] proposed a different strategy for mining HUI by focusing on pattern growth without candidate generation. Their approach reduces computational overhead by avoiding the generation of unnecessary candidates, enhancing efficiency. The algorithm employs a pattern growth strategy to construct HUI directly, eliminating the need for candidate itemset generation. This design choice streamlines the mining process, making it more efficient and suitable for large datasets.

Wu et al. [3] presented UBP-Miner, an efficient bit-based HUI mining algorithm. This algorithm employs a bitmap representation to encode itemsets, significantly reducing memory usage and computational costs. By utilizing bit operations, the algorithm efficiently generates HUI. The integration of a utility-based pruning strategy further enhances its performance. UBP-Miner demonstrates substantial efficiency improvements compared to traditional HUIM algorithms, making it a promising solution for large-scale applications.

Mai et al. [4] propose an algorithm for mining non-redundant high-utility association rules. The paper addresses the challenge of efficiently discovering association rules that maximize utility without redundancy. The proposed algorithm employs an innovative approach to prune the search space and optimize the mining process. In order to mine highly useful patterns that are strongly correlated, Saeed et al. [5] provide an effective technique that is based on utility trees. The paper focuses on improving the mining process by leveraging utility trees and correlation analysis. The algorithm optimizes pattern generation and pruning techniques, resulting in enhanced efficiency and accuracy in pattern discovery. The authors substantiate the algorithm's performance through comprehensive experimentation and comparative analysis.

Dam et al. [6] present the CLS-Miner algorithm, the paper addresses the need for mining closed itemsets with high utility, which is essential for pattern discovery in various applications. The proposed algorithm utilizes efficient data structures and pruning strategies to streamline the mining process, resulting in improved efficiency and scalability. Experimental evaluations demonstrate the effectiveness of the CLS-Miner algorithm in generating high-utility closed itemsets.

In their paper, Wu et al. [7] proposed effective pruning strategies for high-utility itemset mining. They introduced a novel pruning technique that significantly reduces the search space and computational complexity of high-utility itemset mining. Their approach optimizes the mining process, making it more efficient and practical for real-world applications.

Wu and his team [8] extended their work to fuzzy high-utility pattern mining in a parallel and distributed Hadoop framework. This research explores the integration of fuzzy logic with high-utility itemset mining, addressing the uncertainty and vagueness often present in real-world data. Their approach enables efficient processing of large-scale data using distributed computing platforms like Hadoop, making it suitable for big data scenarios.

Wang et al. [9] proposed an improved strategy for high-utility pattern mining algorithms. Their work enhances the efficiency of existing high-utility itemset mining algorithms by introducing new optimization techniques and heuristics. This research focuses on reducing computational overhead while maintaining the quality of the mined high-utility itemsets.

The rapid advancements in technology and the proliferation of big data have revolutionized the field of medical research, allowing for a deeper understanding of the complex interplay between genetic variants and diseases. Kim et al. [10] proposed a medical big data analysis system designed to unearth associations between genetic variants and diseases. The study leverages the vast amounts of data available in the medical domain to enhance the understanding of genetic contributions to various diseases. By employing sophisticated analytical techniques, this system aims to identify meaningful correlations, providing valuable insights for personalized medicine and targeted therapeutic approaches.

In the domain of big data processing, Sethi, Krishan, and Ramesh [11] presented the P-FHM+ algorithm, an innovative approach for mining HUI in large datasets. HUI mining is crucial for numerous applications, including retail, healthcare, and market basket analysis. The P-FHM+ algorithm demonstrates the significance of parallel processing in handling the immense volume of data efficiently. By optimizing the mining process, this algorithm enables the extraction of valuable patterns that have a high utility, contributing to informed decision-making and resource optimization.

Patel, Shah, and Patel [12] proposed an efficient HUIM approach employing predicted utility co-exist pruning. This method represents a step forward in enhancing the efficiency and accuracy of HUIM. By predicting the utility of itemsets and incorporating co-exist pruning strategies, the approach minimizes unnecessary computations, resulting in a more streamlined mining process. The utilization of such predictive techniques is paramount in handling big data efficiently and deriving actionable insights in diverse domains.

Qu, Liu, and Fournier Viger [13] tackle the crucial issue of HUIM without candidate generation in their study. In order to find HUI from a provided dataset without producing superfluous candidates, the authors provide effective techniques to deal with this problem. This paper's method is vital for optimizing resources and lowering computing complexity in mining HUI, which is important for many practical applications like market basket analysis.

Mai and Nguyen [14] propose an efficient approach specifically focused on mining closed HUI and generators. Closed itemsets are valuable as they compactly represent frequent patterns and can significantly reduce the search space. This paper provides insights into an optimized method for mining closed HUI, contributing to efficient pattern discovery in various domains. The presented approach is expected to enhance the performance of HUIM by identifying essential patterns more effectively.

Dawar et al. [15] introduce a hybrid framework designed for mining high-utility itemsets, particularly in sparse transaction databases. Sparse databases pose unique challenges due to the limited availability of data, making traditional mining approaches less effective. The hybrid framework combines different strategies to address these challenges and efficiently mine high-utility itemsets. This paper presents an innovative approach that enhances the applicability of high-utility itemset mining in scenarios where data sparsity is a prevalent concern.

In their work, Mathe John Kenny Kumar and Dipti Rana [16] presented a recent advancement in HUIM known as RSPHUIM. The authors introduced a novel approach to efficiently discover HUI within short time periods. They focused on enhancing the speed and effectiveness of mining HUI, making the approach well-suited for real-time or time-sensitive applications. By addressing the short period aspect, this work contributes to optimizing mining processes in scenarios where the utility of items fluctuates rapidly.

Chen et al. [17] introduced TOPIC (Top-k High-Utility Itemset Discovering), Their work focused on identifying the most valuable itemsets according to the utility measure. By prioritizing and retrieving the top-k high-utility itemsets, the approach assists in decision-making processes for businesses and organizations. The research by Chen et al. contributes to the field by optimizing the search for the most valuable itemsets, aligning with practical needs in various domains.

Furthermore, Han et al. [18] presented an efficient algorithm for top-k HUI mining on massive data. Their work addressed the scalability issue, crucial for handling large datasets effectively. By improving the efficiency of mining HUI on massive data, their approach enables the analysis of vast amounts of information in real-world applications. The research by Han et al. is essential for handling the challenges posed by the increasing volume of data, emphasizing the importance of scalability in HUIM.

Djenouri and Comuzzi [19] present a novel approach that combines Apriori heuristic with bio-inspired algorithms to address the FIM problem. FIM is a fundamental task in data mining, essential for discovering patterns and associations in large datasets. The Apriori algorithm is a classical and widely used method in this domain, but it can be computationally expensive for large datasets due to its multiple database scans. In their work, the authors propose a synergistic integration of the Apriori heuristic with bio-inspired algorithms.

Nguyen et al. [20] focus on a specific variant of FIM, known as mining high-utility itemsets, in dynamic profit databases. High-utility itemsets consider both the frequency of item occurrences and their corresponding utility values. Utility mining is essential in various real-world applications like retail, where not all items have equal importance or profit values. The authors propose an approach to efficiently mine high-utility itemsets in dynamic profit databases, where the profit associated with items may change over time. This dynamic aspect introduces additional complexity to the mining process, necessitating innovative algorithms to adapt to changes and efficiently extract high-utility itemsets. Their proposed method employs a tailored strategy to handle the dynamic nature of profit databases, ensuring the accuracy and relevance of high-utility itemsets over time. The study showcases the effectiveness of their approach through extensive experimentation on diverse datasets, demonstrating its capability to adapt to dynamic changes and provide meaningful insights into high-utility itemsets in evolving environments.

These researchers work demonstrate significant advancements in the field of HUIM, addressing challenges related to candidate generation, closed itemset mining, and handling sparse transaction databases, significance of leveraging big data analysis techniques to enhance our understanding of complex relationships in both medical and HUIM domains, mining closed HUI and generators showcasing innovative techniques to expedite the process and enable practical applications in various domains. Their contributions lay the foundation for efficient and effective HUIM techniques with potential applications across various domains. These papers collectively contribute to the field of high-utility itemset mining by addressing key challenges such as computational complexity, scalability, and the handling of uncertain data. They propose effective pruning strategies, explore fuzzy high-utility mining, and introduce novel optimization techniques to enhance the efficiency and applicability of high-utility itemset mining algorithms.

Table 1. Comparative Analysis of Researchers work

Ref	Algorithms Used	Dataset Used	Advantages	Disadvantages
[1]	Particle filter-based HUIM algorithm	Criteo	Efficient and effective for mining HUI from large datasets	May be sensitive to the selection of hyperparameters

[3]	Bit based HUIM algorithm	Retail	Efficient and scalable for mining HUI from large datasets	May not be effective for mining rare HUI.
[4]	Non-redundant high-utility association rule mining algorithm	Retail	Efficient mining non-redundant high-utility association rules	May not be effective for mining HUI from large datasets
[6]	CLS-Miner	Multiple real-world and synthetic	CLS-Miner is efficient and effective for HUIM, especially for large and complex datasets.	CLS-Miner requires a large amount of memory, especially for large datasets.
[7]	HUIM algorithm with effective pruning strategies.	Multiple real-world and synthetic	The proposed algorithm is efficient and effective for HUIM, especially for large and complex datasets.	The algorithm requires a large amount of memory, especially for large datasets.
[8]	Parallel and distributed Hadoop framework for fuzzy high-utility pattern mining (FHUPM).	Multiple real-world and synthetic	The proposed framework is efficient and scalable for FHUPM on large datasets.	The framework requires a large amount of resources, such as CPU and memory.
[9]	Improved strategy for removing non-candidate items from the global header table and local header table, EFIM algorithm.	Multiple real-world and synthetic	The proposed strategy is effective in improving the efficiency of EFIM for HUIM.	The strategy is only applicable to the EFIM algorithm.
[10]	Associations between genetic variants and diseases.	Medical	The proposed system is effective in discovering high-utility associations between genetic variants and diseases.	The system is complex and requires high computational resources.
[11]	P-FHM+	Synthetic and real-world	Efficient mining in big data, parallel processing	Not handling very large-scale datasets efficiently
[12]	Predicted Utility Co-exist Pruning	Synthetic and real-world	Enhanced efficiency via utility prediction	Complexity in utility prediction
[13]	UEP: Utility-based Early Pruning algorithm, UEP+: Utility-based Early Pruning with prefix utility, and HUP	Synthetic and real-life	Efficient HUIM without candidate generation	Not suitable for very large-scale datasets

[14]	CHUI Miner algorithm	Real-world retail dataset	Focus on closed HUI generators	Limited to retail datasets, potential bias towards closed itemsets
[15]	Hybrid framework that combines the Apriori algorithm with a utility-based pruning strategy	Sparse transaction	Efficient mining in sparse transaction databases	Complexity due to hybrid approach
[17]	Transaction merging, database projection, minUtil threshold raising strategies, array-based utility technique	Retail, synthetic	Efficient, scalable, can handle positive and negative utility values	Requires a large amount of memory, can be slow for large datasets
[18]	Apriori algorithm, utility-list structure	supermarket transactions	Efficient for mining top-k HUI from massive datasets	Not suitable for mining HUI in dynamic datasets
[20]	Utility-list structure	supermarket transactions	Efficient for mining high-utility itemsets in dynamic profit databases	Not suitable for mining high-utility itemsets from very large datasets

III. RELATED WORK

A. Frequent Itemset Mining

Frequent itemset mining (FIM) is a key component of data analysis, enabling the identification of frequent itemsets for decision-making. However, conventional FIM methods overlook valuable datasets with low frequency but high weight.

FIM is a fundamental task in data mining and association rule learning, aimed at discovering recurring patterns within a dataset. This process involves identifying sets of items that frequently co-occur in transactions, databases, or other structured data. The underlying principle is to unveil associations or relationships among items, enabling insights into customer behavior, market basket analysis, and various domains where understanding item co-occurrence is essential. The discovery of frequent itemsets lays the foundation for generating association rules, which are logical implications describing relationships between different items in a transaction. These rules are valuable for decision-making, recommendation systems, and other applications where understanding the inherent associations within data is crucial.

Several algorithms have been developed to efficiently mine frequent itemsets from large datasets. Some prominent algorithms include: Apriori Algorithm, FP-Growth, Eclat, etc. Figure 1 shows the various FIM Algorithms.

The extensive exploration of FPM issues by numerous researchers stems from its rich applications in various data mining tasks such as classification, clustering, and outlier analysis, as elucidated by Aggarwal et al [21]. FPM holds a crucial position in enhancing methods for classifying or clustering data sets and detecting outliers or anomalies. Its pivotal role extends to performing diverse tasks in data mining, contributing significantly to the identification of hidden patterns that recurrently exist in datasets. These discovered patterns are essential for generating association rules employed in data analysis. FPM serves as a foundational step in unveiling frequent patterns within datasets, shaping the landscape of data analysis.

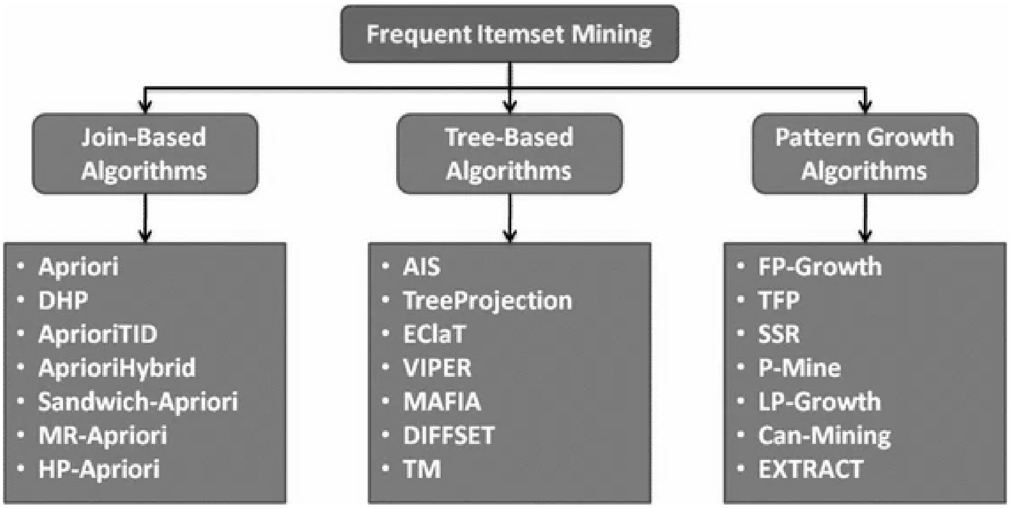

Figure 1. Frequent Itemset Mining Algorithms

Numerous algorithms proposed by various researchers have sought to enhance frequent pattern mining (FPM) techniques. Despite these efforts, there is still a need for further improvements in the performance of existing FPM algorithms, particularly in dealing with vast datasets characterized by an increasingly large volume of data. A significant drawback of many current algorithms is their inadequacy in efficiently mining extensive datasets, as they grapple with prolonged run times and substantial memory consumption. Jamsheela et al. [22] identified two primary challenges faced by most FPM algorithms: extended execution times and extensive memory usage in the pursuit of mining all concealed frequent patterns.

This research endeavors to address these challenges by developing an algorithm capable of efficiently mining significant frequent patterns within datasets, even as data volumes continue to grow. The primary objective is to construct an algorithm that can adapt to the continuous increase in data size over time. This paper aims to review both the advantages and disadvantages of noteworthy and recent FPM algorithms. Through this review, the intention is to pave the way for the development of a more efficient FPM algorithm. Ongoing research efforts persist in refining existing algorithms and exploring novel techniques to overcome the challenges posed by evolving data characteristics and increasing computational demands in the realm of frequent pattern mining. To address this limitation, HUIM has been proposed by researchers, which considers both frequency and weight to identify HUIs.

B. High-utility itemset mining (HUIM)

HUIM algorithms can be categorized into Apriori-based and tree structure-based approaches. The Apriori algorithm iteratively searches the dataset multiple times to filter out frequent itemsets. Similarly, Apriori-based HUIM algorithms generate candidates at each iteration to extract HUIs. To overcome the computational challenges of Apriori, the FP-growth algorithm, based on a tree structure, requires only two scans of the dataset. Tree structure-based HUIM algorithms leverage the FP-growth idea to determine HUIs using conditional tree structures.

For mining HUIs, Yao and Hamilton presented the problem in 2004 [26]. Even though it is an approximation and could miss certain HUIs, they came up with the UMining technique to extract useful item sets. Liu et al. [16] created the Two-Phase approach to address this restriction after realizing it. To simplify the search space and guarantee the extraction of all HUIs, the Two-Phase method incorporates a new upper bound pruning characteristic named TWU (Transaction Weighted Utilization). The TWU algorithm contains two separate steps.

Phase one involves finding potential HUIs that do not have a TWU lower than the minimal utility requirement. In the second stage, it re-scans the database to get the HUIs, and then it evaluates the usefulness of each candidate. On the other hand, the Two-Phase approach has problems with memory and time efficiency, mostly because the first phase generates a large number of candidates.

1. ALGORITHMS

A. High Utility Mining (HUM) Algorithms:

Finding patterns in datasets that show high utility or value is the goal of HUM algorithms. It is the frequency and significance of these patterns that define them. The unique difficulties of HUM have inspired the creation of a number of algorithms tailored to the problem. Here, we discuss some of the notable existing algorithms in this domain:

1. Apriori-based Algorithms:

These algorithms, such as AprioriHU and UP-Growth, extend the classic Apriori algorithm to handle HUI. They adopt a level-wise search strategy to generate candidate itemsets and prune those that do not meet the utility threshold. Apriori-based algorithms are efficient for mining HUI but can suffer from the inherent drawback of generating a large number of candidate itemsets.

2. FP-Growth-based Algorithms:

FP-Growth-based algorithms, including FHM and UP-Growth+, leverage the FP-Growth algorithm for HUM. They construct a frequent pattern tree (FP-tree) to efficiently represent the dataset and generate high utility patterns through a recursive growth process. FP-Growth-based algorithms are known for their ability to handle large datasets and reduce the number of generated patterns compared to Apriori-based algorithms.

3. HUI-Miner:

HUI-Miner is an algorithm specifically designed for mining HUI It adopts a two-phase approach, starting with a candidate generation phase and followed by a utility calculation phase. HUI-Miner utilizes an efficient tree structure, called HUI-Tree, to prune unpromising candidate itemsets early in the process. This algorithm achieves high efficiency by avoiding redundant calculations.

4. UF-growth:

UF-growth is an algorithm designed for mining high utility episodes, which are temporal patterns with associated utilities. It adopts a vertical representation of the dataset and employs a growth strategy to generate high utility episodes. UF-growth utilizes an effective utility-list structure to prune candidate episodes efficiently. This algorithm is particularly suitable for mining high utility episodes in time-series data.

For HUI workloads, these existing algorithms offer valuable solutions. But, the features of the dataset and the mining tasks at hand can cause the performance of the algorithms to differ. Improved efficiency and efficacy of HUIM in many areas are ongoing goals of researchers who are actively exploring and developing new algorithms.

IV. RESULT ANALYSIS

A. Performance Parameters

The most common performance parameters used to evaluate FIM and HUIM are:

- Runtime: The total time required to mine the HUI
- Memory usage: The amount of memory required to mine the FIM and HUI
- Number of HUI mined: The total number of HUI mined
- Accuracy: The percentage of HUIM and FIM that are correctly mined
- Scalability: The ability of the algorithm to handle large and complex datasets

B. Dataset Descriptions:

Dataset are available at Kaggle websites.

1. Market Basket Analysis Data:

- Description: Market basket analysis data is commonly used to analyze customer purchasing patterns in retail or e-commerce settings. It consists of transactional data where each row represents a customer transaction, and the columns represent items or products purchased in that transaction.
- Attributes: The dataset typically includes attributes such as transaction ID, date/time of the transaction, quantity, item price and a list of items purchased in that transaction.

2. Mushroom Dataset:

- Description: The mushroom dataset is often employed for classification tasks, particularly in the domain of edible or poisonous mushroom identification. It contains features that describe various attributes of mushrooms, such as their cap shape, color, odor, habitat, etc.
- Attributes: The mushroom dataset includes attributes like cap shape, cap color, odor, gill attachment, gill color, stalk shape, stalk surface, population, habitat, etc. The target variable indicates whether the mushroom is edible or poisonous.

3. Chess Dataset:

- Description: The chess dataset is designed for analyzing and exploring chess game data, offering insights into the strategies and outcomes of chess matches. It is commonly used for tasks related to move pattern recognition, opening strategies, and game result prediction
- Attributes: Game ID,Event, Site, Date, White Player,Black Player, Black Elo, Moves, Result, Time Control, Ply Count.

V. RESULTS

In our comparative analysis of FIM algorithms and HUIM, focusing on various algorithms used by researchers across multiple datasets, the results provided valuable insights into their respective execution times. Table 1 shows the running time required for 2 datasets namely CHESS and RETAIL and respective graph is shown in figure 2.

Table 2. Time Comparison graph of researcher's algorithms

	DATASET (CHESS) TIME (S)	DATASET (RETAIL) TIME (S)
[1] PF-HUIM	224	325
[2] UP-FPG	980	170
[4] LNR-HNR	140	30
[5] COHUIM	170	47
[8] EFUPM	90	40
[11] P-FHM	130	80
[12] PUCP-Miner	280	150
[14] HUCI- Miner	120	100
[18] TOP-HUI	80	60

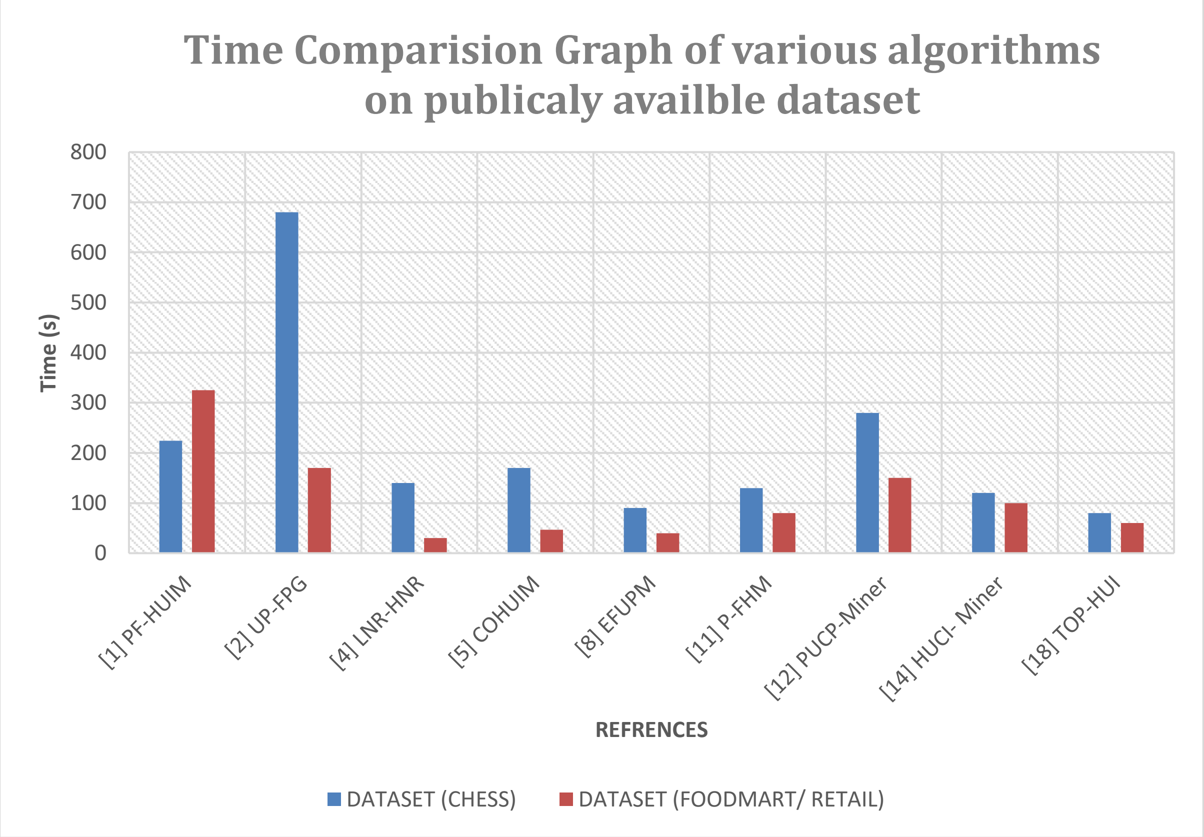

Figure 2. Time Comparison Graph of various algorithms on publically available dataset.

VI. CHALLENGES

FIM and high utility-based mining algorithms, while powerful in extracting patterns from datasets, face several challenges that impact their efficiency and effectiveness. Here are some key challenges associated with these mining algorithms:

- Exponential Growth of Itemsets: As the dataset size increases, the number of potential itemsets grows exponentially. This leads to a combinatorial explosion, making it computationally expensive to discover and enumerate all itemsets.
- Memory Consumption: Itemset mining algorithms often need to store and manage a large number of itemsets in memory during the mining process. This can become a significant challenge for datasets with high dimensionality or when working with limited system memory.
- Scalability with Utility Measures: Defining utility measures that accurately capture the importance of items in real-world applications is challenging. Balancing the need for meaningful utility measures with the scalability of mining algorithms becomes crucial for handling large-scale datasets.
- Multiple Objective Optimization: High utility-based mining involves optimizing multiple conflicting objectives, such as maximizing utility while minimizing costs. Finding a balance between these objectives and designing algorithms that can handle diverse utility functions is a complex task.
- Incremental Mining for Evolving Data: Adapting high utility-based mining to changing data distributions and evolving datasets requires algorithms capable of incremental updates. Ensuring efficiency in updating utility values and discovering patterns in dynamically changing datasets is a non-trivial challenge.
- Mining High-Dimensional Data: Handling high-dimensional data, where items are numerous and diverse, poses a challenge for high utility-based mining. The identification of utility-rich patterns in such datasets demands advanced techniques to navigate the vast search space.
- Interactivity and User Involvement: HUM often involves user-defined utility functions, requiring user input to specify the importance of different items. Ensuring user-friendly interfaces and involving domain experts in the mining process can be challenging but is essential for meaningful results.

Addressing these challenges involves ongoing research efforts to develop more efficient algorithms, improve scalability, and enhance adaptability to diverse data characteristics. Additionally, the integration of machine learning and optimization techniques plays a vital role in advancing the capabilities of FIM and high utility-based mining algorithms.

VII. CONCLUSION

This study performs utility-based and FIM algorithms comparisons, recognizing the imperative need for efficient pattern mining algorithms capable of handling the escalating volumes of data across diverse domains. By conducting a comprehensive review, the dual aspects of efficiency and effectiveness have been emphasized in the context of extracting valuable patterns from large-scale datasets. The significance of utility-based mining lies in its capacity to address real-world scenarios, where items possess varying degrees of importance. Simultaneously, FIM contributes by uncovering recurring patterns, unveiling associations and dependencies within datasets. The synthesis of recent advancements in both utility-based and FIM provides a nuanced understanding of their strengths and limitations. Efficiency considerations, encompassing algorithmic optimizations, parallel computing, and scalability, are crucial for ensuring computational tractability in the face of vast datasets. Meanwhile, effectiveness evaluations scrutinize the quality and relevance of discovered patterns, reinforcing their utility across diverse applications. The overarching theme of this review highlights the delicate balance needed between computational efficiency and the extraction of meaningful patterns. Striking this balance is vital for researchers, practitioners, and decision-makers navigating the challenges of data mining and knowledge discovery. By examining state-of-the-art techniques and methodologies, this review seeks to empower stakeholders with a comprehensive perspective on the current landscape in utility-based and FIM. Ultimately, it aims to provide valuable insights into future directions, fostering advancements that enhance both the efficiency and effectiveness of pattern discovery in the dynamic and expansive realm of data mining.

VIII. REFERENCES

[1]. Yang Yang, Jiaman Ding, Honghai Wang, Huifen Xing, En Li, "A High Utility Itemset Mining Algorithm Based on Particle Filter", Mathematical Problems in Engineering, vol. 2023, Article ID 7941673, 15 pages, 2023. https://doi.org/10.1155/2023/7941673

[2]. Liu, Y.; Wang, L.; Feng, L.; Jin, B. Mining High Utility Itemsets Based on Pattern Growth without Candidate Generation. Mathematics 2021, 9, 35. https://doi.org/10.3390/math9010035

[3]. Peng Wu, Xinzheng Niu, Philippe Fournier-Viger, Cheng Huang, and Bing Wang. 2022. UBP-Miner: An efficient bit based high utility itemset mining algorithm. Know. -Based Syst. 248, C (Jul 2022). https://doi.org/10.1016/j.knosys.2022.108865

[4]. Mai T, Nguyen LTT, Vo B, Yun U, Hong TP. Efficient Algorithm for Mining Non-Redundant High-Utility Association Rules. Sensors (Basel). 2020 Feb 17;20(4):1078. doi: 10.3390/s20041078. PMID: 32079200; PMCID: PMC7070778.

[5]. Rashad Saeed, Azhar Rauf, Fahmi H. Quradaa, Syed Muhammad Asim, "Efficient Utility Tree-Based Algorithm to Mine High Utility Patterns Having Strong Correlation", Complexity, vol. 2021, Article ID 7310137, 18 pages, 2021. https://doi.org/10.1155/2021/7310137

[6]. Dam, TL., Li, K., Fournier-Viger, P. et al. CLS-Miner: efficient and effective closed high-utility itemset mining. Front. Comput. Sci. 13, 357–381 (2019). https://doi.org/10.1007/s11704-016-6245-4

[7]. Jimmy Ming-Tai Wu, Jerry Chun-Wei Lin, and Ashish Tamrakar. 2019. High-Utility Itemset Mining with Effective Pruning Strategies. ACM Trans. Knowl. Discov. Data 13, 6, Article 58 (December 2019), 22 pages. https://doi.org/10.1145/3363571

[8]. Wu, J. M.-T., Srivastava, G., Wei, M., Yun, U., & Lin, J. C.-W. (2021). Fuzzy high-utility pattern mining in parallel and distributed Hadoop framework. Information Sciences, 553, 31-48. 10.1016/j.ins.2020.12.004

[9]. Le Wang, Shui Wang, Haiyan Li, Chunliang Zhou, "Improved Strategy for High-Utility Pattern Mining Algorithm", Mathematical Problems in Engineering, vol. 2020, Article ID 1971805, 11 pages, 2020. https://doi.org/10.1155/2020/1971805

[10]. D. Kim et al., "Medical Big Data Analysis System to Discover Associations between Genetic Variants and Diseases," ICC 2021 - IEEE International Conference on Communications, Montreal, QC, Canada, 2021, pp. 1-6, doi: 10.1109/ICC42927.2021.9500497.

[11]. Sethi, Krishan & Ramesh, Dharavath & Edla, Damodar. (2018). P-FHM+: Parallel high utility itemset mining algorithm for big data processing. Procedia Computer Science. 132. 918-927. 10.1016/j.procs.2018.05.107.

[12]. Patel, S. B., Shah, S. M., & Patel, M. N. (2022). An Efficient High Utility Itemset Mining Approach using Predicted Utility Co-exist Pruning. International Journal of Intelligent Systems and Applications in Engineering, 10(4), 224–230

[13]. Qu, Jun-Feng & Liu, Mengchi & Fournier Viger, Philippe. (2019). Efficient Algorithms for High

Utility Itemset Mining Without Candidate Generation. 10.1007/978-3-030-04921-8_5.

[14]. Mai, Thang & Nguyen, Loan. (2017). An efficient approach for mining closed high utility itemsets and generators. Journal of Information and Telecommunication. 1. 193-207. 10.1080/24751839.2017.1347392.

[15]. Dawar, Siddharth et al. "A hybrid framework for mining high-utility itemsets in a sparse transaction database." Applied Intelligence 47 (2017): 809-827.

[16]. Mathe John Kenny&nbsp;Kumar and Dipti Rana. 2023. RSPHUIM: Recent Short Period High Utility Itemset Mining. SN Comput. Sci. 4, 5 (Sep 2023). https://doi.org/10.1007/s42979-023-01967-y

[17]. Chen, Jiahui & Shicheng, Wan & Gan, Wensheng & Chen, Guoting & Fujita, Hamido. (2021). TOPIC: Top-k High-Utility Itemset Discovering.

[18]. Han, Xixian & Liu, Xianmin & Li, Jianzhong & Gao, Hong. (2020). Efficient Top-k High Utility Itemset Mining on Massive Data. Information Sciences. 557. 10.1016/j.ins.2020.08.028.

[19]. Djenouri, Y.; Comuzzi, M. Combining Apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem. Inf. Sci. 2017, 420, 1–15

[20]. Nguyen, L.T.; Nguyen, P.; Nguyen, T.D.; Vo, B.; Fournier-Viger, P.; Tseng, V.S. Mining high-utility itemsets in dynamic profit databases. Knowl.-Based Syst. 2019, 175, 130–144

[21]. Aggarwal CC (2014) An introduction to Frequent Pattern Mining. In: Aggarwal CC, Han J (eds) Frequent Pattern Mining. Springer, Basel, pp 1–14

[22]. Jamsheela O, Raju G (2015) Frequent itemset mining algorithms: a literature survey. In: Paper presented at the 2015 IEEE international advance computing conference (IACC), Banglore.

Cite this article as :

Nishigandha Mhatre, Srijita Bhattacharjee, "Efficiency and Effectiveness in Utility-Based and Frequent Itemset Mining : A Comprehensive Review", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 11 Issue 1, pp. 179-192, January-February 2024.

Journal URL : https://ijsrset.com/IJSRSET2310670