Review Article

International Journal of Advanced Computer Research, Vol 13(62)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2023.1362001

An analysis and literature review of algorithms for frequent itemset mining

Mrinabh Kumar[1*], **Animesh Kumar Dubey**[2]
Department of Computer Science, Patel College of Science & Technology Bhopal, Madhya Pradesh, India[1]
Assistant professor, Department of Computer Science, Patel College of Science & Technology Bhopal, Madhya Pradesh, India[2]

Received: 10-January-2023; Revised: 15-March-2023; Accepted: 30-March-2023

Abstract

The data mining process should be led by domain knowledge. It includes different aspects including the selection of the data, interpretation, extraction, and transformation. In this paper different domains have been covered for the analysis of various data mining algorithms. The main emphasis on the algorithms which are mainly used for the extraction and discovering of interesting patterns and relationships. Various data mining algorithms, such as sequential pattern discovery using equivalence classes (SPADE), k-means, Apriori algorithm, FP-Growth and others, were discussed in this paper. The reviews and analysis of the advantages and disadvantages of various data mining approaches have been explored with advantages and limitations. In summary, this paper provides a comprehensive understanding of data mining approaches and their potential applications in various fields.

Keywords

Data mining, Domain knowledge, Preprocessing, Knowledge discovery.

1.Introduction

In almost every domain, data mining algorithms are utilized to uncover meaningful information and patterns. These algorithms have become a popular tool for analyzing and interpreting large datasets, allowing businesses, engineers, and medical professionals to make informed decisions [1–3]. In the fields of engineering, medicine, and business, data mining algorithms have proven to be particularly useful. They are used to extract insights from data generated by various systems, including sensors, electronic medical records, and customer databases [4, 5].

Data mining algorithms have been instrumental in engineering applications, such as predicting equipment failure and optimizing manufacturing processes [6–10]. In medicine, these algorithms have been used to identify patterns in patient data, helping to diagnose diseases and develop treatment plans [11, 12]. In business, data mining algorithms are used to analyze customer behavior, identify trends, and make strategic decisions [13, 14].

The task of high-utility data mining is a crucial area in the field of knowledge discovery. Traditional frequent pattern mining algorithms have been enhanced by incorporating the methodology presented in [13], which allows for the identification of groups of frequently occurring items. However, frequent pattern mining techniques are limited in their ability to recognize the relative importance of individual items and assume that every item has the same value and occurs only once per transaction [13–17].

To address these limitations, high-utility pattern mining has emerged as a powerful approach that considers the weight of individual items and their contribution to the overall usefulness of a set. By considering the value of each item, high-utility pattern mining enables the extraction of more meaningful and valuable insights from value-based databases. This methodology has found numerous applications in different application areas including business, finance and market basket implications etc. [18–20].

The main aim of this paper is to review and analyses data mining approaches and discuss the advantages and disadvantages of the approaches. Data mining

[*] Author for correspondence

approaches can be applicable in different fields including education, healthcare, retail, business analytics, finance, economy, and growth, etc. (*Figure 1*).

This paper comprises several sections, including an introduction, a literature review in section 2, a discussion and analysis in section 3, and a conclusion in section 4.

Education
Healthcare
Finanace
Data Mining
Retail
Economy and growth
Business analytisc

Figure 1 Data mining applicability in different domains

2.Literature review

In this section different related worked have been discussed and analyzed.

The PNPFI algorithm was proposed by Zhang et al. [21] to perform parallel frequent itemset mining using the N-list structure, P-Subsume, and a load balancing strategy. The algorithm's effectiveness was evaluated through experiments, which showed that PNPFI outperformed existing algorithms, achieving a performance improvement of up to 79% and reducing memory usage by up to 90%. While the approach is well-suited for dense datasets with numerous P-Subsume, its effectiveness may vary depending on the dataset.

Agarwal et al. [22] proposed an algorithm for mining high average-utility itemsets (HAUIs) considered both the length of itemsets and their utilities, providing a more accurate representation of real-world scenarios. The traditional method for finding high-utility itemsets (HUIs) calculates the individual utility of an itemset, neglecting the length of itemsets. The algorithm is demonstrated using a numerical example, and experiments conducted on real-life databases show that it efficiently discovers the entire set of HAUIs. While the proposed approach can provide better decision-making support in different fields, future research should explore the algorithm's limitations in terms of scalability and complexity.

The FIMIU algorithm was proposed as a solution to the challenge of using market basket analysis [23]. It has been used for behaviour analysis and the environment considered was data streaming [23]. Experimental results show that FIMIU is memory efficient, but requires slightly more time to run. The advantages of the algorithm include its ability to analyze a larger portion of the data stream, making it more suitable for real-time decision-making. However, the algorithm's limitations should be further investigated in future research, such as its scalability and potential trade-offs between memory usage and runtime.

Extended partitioning algorithm (EPA) was proposed to reduce the number of additional scans required for mining quantitative association rules from multidimensional agricultural datasets [24]. The EPA method is compared with the traditional partitioning algorithm (PA) and found to be more efficient in terms of reducing the number of additional scans required for aggregation of multidimensional attributes at different levels. The proposed EPA method successfully reduces the time complexity of mining quantitative association rules and provides better performance. However, the limitation of the EPA method is that it may not be suitable for datasets with irregular or non-uniform distributions of attributes.

A closed itemset property-based multi-objective evolutionary algorithm (CP-MOEA) was proposed by Cao et al. [25]. It was proposed for the frequent set mining and analysis considering transactional databases. Advantages of the proposed algorithm include an improved quality of the mined itemsets, the ability to handle many itemsets, and scalability to large datasets. However, limitations of the algorithm include a high computational cost and the requirement of domain expertise to select appropriate parameters for the algorithm.

High-quality pattern mining (HQPM) was proposed for the problem based on multi objective [26]. They have used an improved version for the objective problem. The results show the effectiveness of the approach considering different occurring patterns for the completeness of the system. It has been considered with different aspects including speed and quality.

In this study [27], a graph-based approach for frequent itemsets mining (FIs) was proposed to efficiently extract useful information from large datasets. The proposed approach represents the complete transactional database as a graph, enabling the storage of all relevant information needed to extract FIs in a single pass. An algorithm that extracts FIs from the graph-based structure is also presented. The advantages of the proposed approach include its ability to handle large datasets in a single pass, reducing the number of passes required for FIs mining, and its scalability. However, the limitations include the need for a domain expert to determine the optimal threshold for frequency, which could affect the quality of the mined frequent itemsets.

Hong et al. [28] proposed a mining approach based on erasable-itemset algorithm. It has been based on the execution time. is a useful pattern extraction method for analyzing production planning in a factory. Several efficient erasable-itemset mining approaches have been developed since its introduction in 2009, but they require a considerable amount of execution time for large product data. The advantages of this approach include its ability to reduce scans of a database and its suitability for analyzing large product data.

Botm-mine was proposed for the mining frequent itemset in the environment of uncertain data flow [29]. It has been experimented on 1 and 2 itemsets. It is found that this approach is efficient in terms of space and computing complexity. However, the limitations of the proposed approach include its reliance on the assumption that data flows are uncertain and the need for further validation on larger datasets.

Nalousi et al. [30] proposed a frequent itemset manning. It is based on weighted subtrees. It is based on FP-Growth algorithm. It has found to be efficient in terms on different weighted datasets. It is found to be efficient in different weighted transactions. *Table 1* shows the review discussion on the latest papers along with the method, ad advantages and limitations.

Table 1 Review discussion on the latest papers along with the method, ad vantages and limitations

S. No	References	Method	Advantages	Limitations
1	[30]	High-utility itemset mining	It is practical in commercial applications since it considers profit in addition to frequency. It can be used to discover high-profit patterns in transactional databases, which can be useful for market basket analysis, product recommendation, and customer segmentation.	The exponential search space, which makes it computationally expensive.
2	[31]	Dynamic prefix tree	Efficiency and memory usage is the main advantage of Dynamic prefix tree.	One limitation of DPT is that it requires significant computational resources, such as memory, to store and manipulate the prefix tree. Another limitation is that it may not be suitable for very large datasets or datasets with a high number of unique items, as the

S. No	References	Method	Advantages	Limitations
				prefix tree can become too large to handle efficiently.
3	[32]	Fast Incremental Updating Frequent Pattern growth algorithm (FIUFP-Growth)	The proposed approach is more efficient than previous methods for incremental association rule mining.	It is unclear how the method would perform on very large or complex databases.
4	[33]	Non-recursive serial FP algorithm, NRFP-growth	The advantages of the improved algorithms are the reduced time and space complexity compared to the original FP-growth algorithm, and the increased speedup ratio and scalability of the parallel algorithm.	The limitations of the algorithms are that they may not be suitable for very large datasets, and the performance improvement may not be significant for datasets with low sparsity. Also, the GPFP-growth algorithm requires access to a GPU, which may not be available for all users.
5	[34]	SMOTE, ROSE, and ADASYN	The advantages of this method are that it improved the creation of association rules and efficiently modeled the bacteria that cause bacterial vaginosis.	The limitations of this study are that it used a preconstructed dataset and was conducted on a specific population, so the results may not be generalizable to other datasets or populations.
6	[35]	Machine learning algorithms	The advantages of this approach are the potential for earlier detection and appropriate therapy for chronic kidney disease, which can slow or halt its progression to an end-stage requiring dialysis or surgery.	The interpretation of the results may be challenging, and there may be ethical considerations regarding the use of personal health data for research purposes.
7	[36]	ck-FARM	The algorithm has been implemented as an R package, making it easy to use and integrate into existing data analysis workflows.	It may be computationally expensive for large datasets, especially when building and adapting fuzzy membership functions.

3.Discussion and analysis

The literature analysis discusses various algorithms proposed for frequent itemset mining. Zhang et al. [21] proposed the PNPFI algorithm for parallel frequent itemset mining, which showed a performance improvement of up to 79% and reduced memory usage by up to 90%. Agarwal et al. [22] proposed an algorithm for mining high average-utility itemsets, providing a more accurate representation of real-world scenarios. FIMIU was proposed to estimate consumer behaviour and demand function more realistically in a data streaming environment [23]. EPA was proposed to reduce the number of additional scans required for mining quantitative association rules from multidimensional agricultural datasets [24]. CP-MOEA and MOEA-PM were proposed for mining frequent and high utility itemsets from transactional databases and high-quality pattern mining, respectively [25]. A graph-based approach was proposed for frequent itemsets mining, which outperformed existing methods in terms of time [27]. Hong et al. [28] proposed a bitmap representation for itemsets in an erasable-itemset mining algorithm to speed up execution.

Various analysis and approaches in data mining have been discussed in the literature. High-utility itemset mining was presented as a practical approach for discovering high-profit patterns in transactional databases, but its main challenge is the exponential search space, which makes it computationally expensive. The dynamic prefix tree approach was suggested as an efficient solution to this problem, although it may require significant computational resources to manipulate the prefix tree [31]. Another approach is for incremental association rule mining, which reduces unnecessary sub-tree construction and rescans of the original database [32]. However, the performance of this approach on very large or complex databases is unclear. Improved algorithms for FP-growth are also discussed, which reduce time and space complexity compared to the original algorithm and improve the speedup ratio and scalability of the parallel algorithm. However, these algorithms may not be suitable for very large datasets or datasets with low sparsity. The GPFP-growth algorithm also requires access to a GPU, which may limit its availability to some users.

The use of machine learning and predictive modeling is suggested as an approach for identifying chronic kidney disease and improving the accuracy of predictions [35]. However, machine learning algorithms require large amounts of data to be trained, and the quality of the data can affect the accuracy of the predictions. Finally, an algorithm is discussed that has been implemented as an R package, making it easy to use and integrate into existing data analysis workflows [36]. However, this algorithm may be computationally expensive for large datasets, especially when building and adapting fuzzy membership functions. Overall, the text presents a range of practical approaches to data mining and machine learning, each with its own advantages and limitations.

The strengths and limitations of each algorithm were evaluated through experimental results. While most algorithms showed significant performance improvements and advantages in terms of scalability, and efficiency. The limitations included requirements for domain expertise to select appropriate parameters, scalability issues, and limitations in handling irregular or non-uniform distributions of attributes. Future research should focus on addressing these limitations and exploring the algorithms' effectiveness in different datasets and real-world scenarios.

4.Conclusion

The literature analysis highlights several algorithms proposed for frequent itemset mining, each with its own advantages and limitations. The algorithms have shown significant improvements in terms of efficiency, scalability, and quality of mined itemsets. However, limitations include the need for domain expertise to select appropriate parameters, scalability issues, and limitations in handling irregular or non-uniform distributions of attributes. Future research should focus on addressing these limitations and exploring the algorithms' effectiveness in different datasets and real-world scenarios. Overall, the previous approaches provide a valuable insight into practical approaches for data mining and machine learning, which can aid in developing more accurate predictive models and improving decision-making processes in various domains.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Shabtay L, Fournier-Viger P, Yaari R, Dattner I. A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data. Information Sciences. 2021; 553:353-75.

[2] Shawkat M, Badawi M, El-ghamrawy S, Arnous R, El-desoky A. An optimized FP-growth algorithm for discovery of association rules. The Journal of Supercomputing. 2022:1-28.

[3] Ghosh M, Roy A, Sil P, Mondal KC. Frequent itemset mining using FP-tree: a CLA-based approach and its extended application in biodiversity data. Innovations in Systems and Software Engineering. 2022:1-9.

[4] Zhang X, Tang Y, Liu Q, Liu G, Ning X, Chen J. A fault analysis method based on association rule mining for distribution terminal unit. Applied Sciences. 2021; 11(11):5221.

[5] Dubey AK, Shandilya SK. A novel J2ME service for mining incremental patterns in mobile computing. In information and communication technologies: international conference, ICT 2010, Kochi, Kerala, India, Proceedings 2010 (pp. 157-64). Springer Berlin Heidelberg.

[6] Happawana KA, Diamond BJ. Association rule learning in neuropsychological data analysis for Alzheimer’s disease. Journal of Neuropsychology. 2022; 16(1):116-30.

[7] Alcan D, Ozdemir K, Ozkan B, Mucan AY, Ozcan T. A comparative analysis of Apriori and FP-growth algorithms for market basket analysis using multi-level association rule mining. In industrial engineering in the Covid-19 Era: selected papers from the hybrid global joint conference on industrial engineering and its application areas, GJCIE 2022, October 29-30, 2022 2023 (pp. 128-37). Cham: Springer Nature Switzerland.

[8] Shahin M, Inoubli W, Shah SA, Yahia SB, Draheim D. Distributed scalable association rule mining over covid-19 data. In future data and security engineering: 8th international conference, FDSE 2021, Virtual Event, 2021, Proceedings 2021 (pp. 39-52). Cham: Springer International Publishing.

[9] Dubey AK, Shandilya SK. Exploiting need of data mining services in mobile computing environments. In international conference on computational intelligence and communication networks 2010 (pp. 409-14). IEEE.

[10] Dubey AK, Gupta U, Jain S. Computational measure of cancer using data mining and optimization. In sustainable communication networks and application 2019 (pp. 626-32). Springer International Publishing.

[11] Ghafoor N, Ahmad M. Nazish Ghafoor, Mansoor ahmad prioritizing effectiveness of algorithms of

association rule mining. Journal of Computational Learning Strategies & Practices. 2021; 1(1):18-30.

[12] Fernandez-Basso C, Ruiz MD, Martin-Bautista MJ. New spark solutions for distributed frequent itemset and association rule mining algorithms. Cluster Computing. 2023:1-8.

[13] Makkar K, Kumar P, Poriye M, Aggarwal S. Improvisation in opinion mining using data preprocessing techniques based on consumer's review. International Journal of Advanced Technology and Engineering Exploration. 2023; 10(99):257-77.

[14] Dubey AK, Kapoor D, Kashyap V. A review on performance analysis of data mining methods in IoT. International Journal of Advanced Technology and Engineering Exploration. 2020; 7(73):193-200.

[15] Suhandi N, Gustriansyah R. Marketing strategy using frequent pattern growth. Journal of Computer Networks, Architecture and High Performance Computing. 2021; 3(2):194-201.

[16] Saxena A, Rajpoot V. A comparative analysis of association rule mining algorithms. In IOP conference series: materials science and engineering 2021 (pp. 1-11). IOP Publishing.

[17] Babu MV, Sreedevi M. Performance analysis on advances in frequent pattern growth algorithm. In 2022 international conference on advances in computing, communication and applied informatics 2022 (pp. 1-5). IEEE.

[18] Anupama CG, Lakshmi C. Approaches to parallelise Eclat algorithm and analysing its performance for K length prefix-based equivalence classes. International Journal of Business Intelligence and Data Mining. 2023; 22(1-2):34-48.

[19] Nikitin E, Kashevnik A, Shilov N. Shopping basket analisys for mining equipment: comparison and evaluation of modern methods. In 2022 31st conference of open innovations association 2022 (pp. 207-13). IEEE.

[20] Yogasini M, Prathibha BN. Comparative analysis on frequent Itemset mining algorithms in vertically partitioned cloud data. In futuristic communication and network technologies: select proceedings of VICFCNT 2020 (pp. 395-402). Springer Singapore.

[21] Zhang F, Zhang Y, Liao X, Jin H. PNPFI: an efficient parallel frequent itemsets mining algorithm. In 22nd international conference on computer supported cooperative work in design 2018 (pp. 172-7). IEEE.

[22] Agarwal R, Gautam A, Saksena AK, Rai A, Karatangi SV. Method for mining frequent item sets considering average utility. In international conference on emerging smart computing and informatics 2021 (pp. 275-8). IEEE.

[23] Amballoor RG, Naik SB. Utility-based frequent itemsets in data streams using sliding window. In international conference on computing, communication, and intelligent systems 2021 (pp. 108-12). IEEE.

[24] Bhatia J, Gupta A. Association rule mining by discretization of agricultural data using extended partitioning algorithm. In 6th international conference for convergence in technology 2021(pp. 1-6). IEEE.

[25] Cao H, Yang S, Wang Q, Wang Q, Zhang L. A closed itemset property based multi-objective evolutionary approach for mining frequent and high utility itemsets. In congress on evolutionary computation 2019 (pp. 3356-63). IEEE.

[26] Fang W, Zhang Q, Sun J, Wu X. Mining high quality patterns using multi-objective evolutionary algorithm. IEEE Transactions on Knowledge and Data Engineering. 2020; 34(8):3883-98.

[27] Halim Z, Ali O, Khan MG. On the efficient representation of datasets as graphs to mine maximal frequent itemsets. IEEE Transactions on Knowledge and Data Engineering. 2019; 33(4):1674-91.

[28] Hong TP, Huang WM, Lan GC, Chiang MC, Lin JC. A bitmap approach for mining erasable itemsets. IEEE Access. 2021; 9:106029-38.

[29] Junrui Y, Jingyi Y. Frequent itemsets mining algorithm for uncertain data streams based on triangular matrix. In international conference on power electronics, computer applications 2021 (pp. 327-30). IEEE.

[30] Nalousi S, Farhang Y, Sangar AB. Weighted frequent itemset mining using weighted subtrees: WST-WFIM. IEEE Canadian Journal of Electrical and Computer Engineering. 2021; 44(2):206-15.

[31] Qu JF, Hang B, Wu Z, Wu Z, Gu Q, Tang B. Efficient mining of frequent itemsets using only one dynamic prefix tree. IEEE Access. 2020; 8:183722-35.

[32] Thurachon W, Kreesuradej W. Incremental association rule mining with a fast incremental updating frequent pattern growth algorithm. IEEE Access. 2021; 9:55726-41.

[33] Wu C, Jiang H. Research on parallelization of frequent itemsets mining algorithm. In 6th international conference on cloud computing and big data analytics 2021 (pp. 210-215). IEEE.

[34] De la Cruz-Ruiz F, Canul-Reich J, Rivera-López R, De la Cruz-Hernández E. Impact of data balancing a multiclass dataset before the creation of association rules to study bacterial vaginosis. Intelligent Medicine. 2023.

[35] Islam MA, Majumder MZ, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. Journal of Pathology Informatics. 2023.

[36] Ho GT, Tsang YP, Wu Q, Tang V. Ck-FARM: an R package to discover big data associations for business intelligence. SoftwareX. 2023.

Mrinabh Kumar is currently pursuing M.Tech degree in Computer Science and Engineering from Patel College of Science and Technology, located in Bhopal, Madhya Pradesh. The college is affiliated with Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV), also located in Bhopal. Mrinabh has completed a Bachelor of Engineering (B.E) degree in Information Technology from LNCT&S, also located in Bhopal, which is affiliated with RGPV BHOPAL. Mrinabh's primary research interests lie in the fields of Data Mining, Optimization, Machine Learning, and Artificial Intelligence.
Email: mrinabhkumar@gmail.com

Animesh Kumar Dubey is an Assistant Professor in the Computer Science and Engineering department at Patel College of Science and Technology in Bhopal, Madhya Pradesh, India. He completed his Bachelor of Engineering (B.E.) and M.Tech degree in Computer Science and Engineering from Rajeev Gandhi Technical University, Bhopal (MP). He has more than 15 publications in reputed peer-reviewed national and international journals and conferences. His research interests include Data Mining, Optimization, Machine Learning, Cloud Computing, and Artificial Intelligence.
Email: animeshdubey123@gmail.com