ISSN 2349-0292
Impact Factor 3.802

GLOBAL JOURNAL OF ADVANCED ENGINEERING TECHNOLOGIES AND SCIENCES

A NOVEL ALGORITHM FOR EFFICIENT FREQUENT PATTERN MINING

Surya Kant Mishra[1], Arpit Solanki[2]

[1]Student, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India
[2]Assistant Professor, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India

ABSTRACT

The rapid expansion of digital data across various sectors—including retail, healthcare, education, and finance—has increased the need for effective techniques that can extract meaningful knowledge from large databases. Among these techniques, association rule mining plays a crucial role in identifying frequent itemsets and uncovering hidden relationships within transactional data. Although traditional algorithms such as Apriori, FP-Growth, and Eclat are widely used for this purpose, their performance depends heavily on a user-defined minimum support threshold. Selecting this value without proper domain understanding often leads to inappropriate outputs, either by generating a large number of insignificant patterns or by overlooking valuable associations.To overcome these limitations, this research introduces a new association rule–based mining algorithm that determines the minimum support threshold mathematically instead of relying on user input. This automated approach simplifies the mining process, enhances accuracy, and reduces the need for expert intervention. The algorithm is implemented using Python and tested on standard datasets obtained from the UCI Repository. Its effectiveness is evaluated by comparing the number of frequent itemsets generated and the execution time with those of the conventional Apriori algorithm. The experimental analysis reveals that the proposed method consistently outperforms Apriori. It produces fewer redundant frequent itemsets, resulting in lower memory usage and improved clarity in the discovered patterns. Additionally, the algorithm demonstrates faster execution times across datasets of varying sizes, highlighting its efficiency and scalability. By eliminating manual threshold selection and improving computational performance, the proposed approach offers a more reliable, practical, and user-friendly solution for frequent pattern mining. This contributes to more intelligent and automated decision-support systems capable of handling large and complex datasets.

KEYWORDS: Data Mining, frequent itemsets, Apriori algorithm.

INTRODUCTION

With the fast progress of information technology, organisations in a variety of industries including banking, transportation, manufacturing, engineering, agriculture, and healthcare are collecting and storing massive volumes of data. Efficiently organising, structuring, and using this vast amount of information has become a critical task. These large datasets have enormous potential, providing useful insights that help with decision-making, query processing, and other critical activities. Data mining is critical for realising this promise. It entails examining massive databases to identify relevant patterns, trends, and linkages. Data mining has become a major study subject and a critical component of modern database systems due to its capacity to extract usable knowledge from huge databases.

The main objective of this research is to improve the quality and reliability of association rule mining by overcoming the limitations associated with manually selected minimum support values. To achieve this, the study first focuses on conducting a comprehensive review of existing association rule mining algorithms, such as Apriori and other frequent pattern mining techniques, to understand their working principles, strengths, and weaknesses in extracting frequent itemsets. Based on the gaps identified, the research aims to design and develop a novel association rule mining algorithm that incorporates a mathematical approach for automatically determining the optimal minimum support threshold, ensuring that the selection of this crucial parameter does not depend on subjective user choices. This automatic calculation is expected to prevent the generation of too many irrelevant patterns due to a low threshold or the loss of significant information due to an excessively high threshold. Finally, the proposed algorithm will be implemented on real-world or benchmark datasets to evaluate its performance, accuracy, and ability to generate meaningful frequent itemsets. Through experimental validation, the study seeks to demonstrate that the newly developed algorithm not only simplifies the mining process by removing the need for user-defined support values but also enhances the overall effectiveness of frequent pattern extraction.

ISSN 2349-0292
Impact Factor 3.802

RELATED STUDY

This section outlines, some noteworthy contributions has been reported :
k. Gulzar et al. (2023) this paper describes a web-based Hospital Information Management System (HIMS) that uses the Apriori data mining method to find common trends in eye problem data, notably myopia, among more than 1000 Chinese university students. The technology improves healthcare decision-making by identifying hidden patterns in patient history, resulting in better diagnosis and service quality. The findings revealed lifestyle correlations to myopia, underlining the potential of data mining in illness prediction. Future work intends to automate HIMS for broader medical applications and to help practitioners detect illness causes through lifestyle analysis.

S. Ding et al. (2022) this paper examines multiple Sequential Pattern Mining (SPM) methods utilising large-scale taxi trajectory data from Beijing, with an emphasis on runtime, memory utilisation, and pattern quality. It demonstrates how trajectory data must be discretised using grid-based segmentation before mining. Experiments show that contiguous-constraint algorithms generate compact, understandable patterns and perform best at low support thresholds, particularly in traffic-related applications. The findings provide a useful guidance for selecting appropriate SPM algorithms for trajectory-based analyses.

S. Nasreen et al (2014) Pattern recognition remains a significant challenge in data mining and knowledge discovery. In this paper, we conduct a comparative analysis of several widely used algorithms designed to extract frequent patterns from large transactional databases. The study evaluates the Apriori algorithm, FP-Growth, Rapid Association Rule Mining (RARM), ECLAT, and the ASPMS algorithm for sensor data streams. Each algorithm is examined in terms of its ability to identify frequent itemsets, along with its strengths, limitations, and suitability for handling large-scale datasets. The findings provide insight into how these methods perform across different data environments and highlight their respective advantages and constraints.

M. Sornalakshmi et al (2021) The study introduces EPDA, an extended parallel and distributed Apriori method based on the Hadoop MapReduce architecture for efficiently mining common patterns from large-scale healthcare datasets. EPDA minimises computation time and communication overhead by spreading data across mappers and trimming infrequent itemsets early on. Experimental results reveal that EPDA beats traditional and existing Apriori variations in terms of speed and rule generation. This makes it suited for detecting hidden trends in healthcare data, allowing for faster and more accurate decision-making.

H. B. Wang et al. (2021) use parallelisation to enhance the Apriori method based on the MapReduce paradigm, aiming to address the classic Apriori algorithm's performance bottleneck when the data set is somewhat larger. After computing the local frequent items on each cluster sub-node, all of them are combined into the global candidate items, and the frequent things that match the requirements are filtered using the minimal support threshold. The new strategy is more efficient since it only scans the transaction database twice and calculates the frequent item set at the same time.

Jie and Gang et al. (2019) introduced an approach for detecting negative association rules in a dataset. Their technique use the Lift measure to detect negative links between distinct item sets, whereas existing algorithms normally only consider positive associations. A lift value more than 1 implies a positive relationship, whilst a number less than 1 indicates a negative one. One disadvantage of the algorithm is its reliance on user-specified thresholds, as well as the lack of certainty that the negative rules created are genuinely interesting or significant.
Kong, H et al. (2018) created a mathematical technique for identifying negative association rules using both common and rare item sets. The study offers several measurements with the goal of developing significant positive and negative association rules. These negative principles can provide important insights that aid decision-making in actual circumstances.

PROPOSED WORK

In a big transactional dataset, such as a retailer database, it is usual for numerous things to be sold or purchased at the same time, hence the database will almost certainly include transactions containing the same set of items. Thus, by taking advantage of these transactions, attempting to discover the frequent patterns or item sets. The proposed research presents a novel method for discovering common item sets that uses a mathematical approach rather than traditional data-mining strategies. This reserch introduces a new algorithm that uses mathematical

method to automatically determine the minimum threshold value needed for frequent item set generation. This removes the need for manual parameter adjustment, which is sometimes a difficult and time-consuming operation for users. By include mathematical computing in the threshold setting process, the approach improves the accuracy and efficiency of frequent pattern detection. Furthermore, the proposed method considerably increases the usability of data mining tools by lowering the technical hurdles that prohibit non-expert users from efficiently applying these techniques. As a result, people with no formal training or specialised understanding in data mining may undertake difficult analyses with greater ease, making the process more accessible, user-friendly, and efficient for a wider range of applications.

Input: Data set of transaction.

Output: any or all of the frequent item sets in Data set.

1: Determine the Minimum Support Threshold, which is equal to Σ ((support (i))/n), where n is the number of items.

2. After a single database scan, add each item to the structure by counting its occurrence in the transactional database. Put them in separate addresses.

3. Scan from the largest item sets in the arrangement to find all the often occurring item sets. Let's say it is a k-item collection. In the event that there are several k-item sets, we evaluate each subsequent k-item set until all of the maximum frequent item sets are obtained.

4(a): The Apriori property states that if the k-item set is frequent, then its subsets are also frequent. If not already present, add the Set and all of its subsets to the frequent-item set table. Proceed to step 7.
4(b): Create subsets of the (k-1)-item set if the k-item set is not common.

5. Instead of scanning the database, use the arrangement to get the support count for any (k-1)-item set subsets that are not included in the frequent item set table.

6: Use min_sup to compare each support count. Proceed to step 4 (a) if the item set is frequent. Proceed to step 4 (b) if the item set is not frequent.

7: Proceed to the structure's address-1, and halt if the address is less than zero.

8: Determine the support count for each item set by scanning them to see whether they are not included in the frequent item set table.

9: Use the min_sup to compare each support count. Proceed to step 4 (a) if the item set is frequent. Proceed to step 4 (b) if the item set is not frequent.

RESULTS ANALYSIS

In this experiment, the goal is to expand the number of items in the dataset. The results highlight the comparative performance of our proposed algorithm against the Apriori algorithm.

ISSN 2349-0292
Impact Factor 3.802

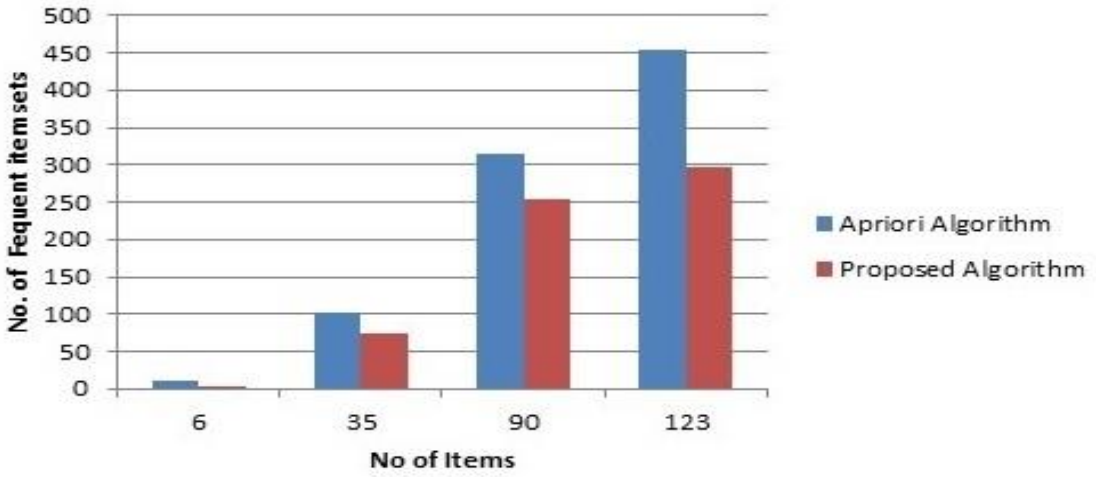

Fig 4.1 Number of Frequent Itemsets

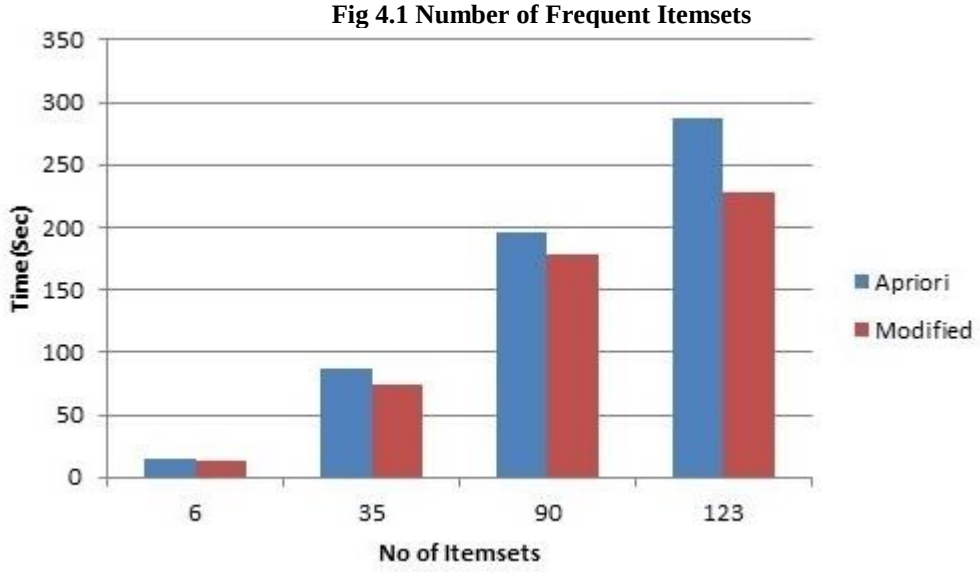

Fig 4.2 Execution Time

CONCLUSION

The study presented in this work aimed to improve the efficiency and accuracy of association rule mining by removing the need for user-defined minimum support threshold values. Traditional algorithms like as Apriori, FP-Growth, and Eclat have demonstrated great success in identifying frequent itemsets; nevertheless, their performance and accuracy are heavily controlled by the value of the minimal support parameter, which frequently necessitates domain knowledge and manual adjustment. To overcome this limitation, a novel association rule-based mining algorithm was introduced that computes the minimum support threshold using a mathematical approach. This eliminates user intervention and enhances the adaptability of the algorithm to different datasets. The proposed algorithm was implemented in Python and evaluated on benchmark datasets from the UCI Repository, with its performance compared to the traditional Apriori algorithm. Experimental results demonstrated that the proposed algorithm consistently outperforms Apriori in terms of execution time and memory efficiency. The algorithm generated fewer redundant frequent itemsets, which reduces storage requirements and simplifies the interpretation of results. Moreover, the automated calculation of the minimum support threshold ensures better rule quality and consistency across diverse datasets without relying on user-specified parameters.

Overall, the study concludes that the proposed algorithm is more efficient, scalable, and user-friendly than the conventional Apriori algorithm. It addresses the key challenge of threshold selection and significantly improves the quality of mined association rules, thereby contributing a more practical and intelligent solution for large-scale data mining applications.

REFERENCES

[1] K. Gulzar, M.A. Memon, S.M. Mohsin, S. Aslam, S.M.A. Akber, M.A. Nadeem "An Efficient Healthcare Data Mining Approach Using Apriori Algorithm: A Case Study of Eye Disorders in Young Adults", mdpi Information 2023, 14, 203. https://doi.org/10.3390/info14040203
[2] S. Ding, Z Li, K. Zhang, F. Mao "A Comparative Study of Frequent Pattern Mining with Trajectory Data", mdpi Sensors 2022, 22, 7608. https://doi.org/10.3390/s22197608
[3] S. Nasreen, M.A. Azam, K. Sehzad, U. Naeem, M.A. Ghazanfar "Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey", Procedia Computer Science 37 (2014) 109 – 116
[4] M. Sornalakshm, S. Balamurali "An efficient apriori algorithm for frequent pattern mining using mapreduce in healthcare data", Bulletin of Electrical Engineering and Informatics, ISSN: 2302-9285, Vol. 10, No. 1, February 2021, pp. 390~403
[5] H. B. Wang, Y. J. Gao, "Research on parallelization of Apriori algorithm in association rule mining", Procedia Computer Science 183 (2021) 641–647
[6] Jie, Z. and Gang, W., 2019, 'Intelligence Data Mining Based on Improved Apriori Algorithm', Journal of Computers, 14(1), pp. 52-62.
[7] Kong, H., An, D. and Ri, J., 2018,'Itemsets of interest for negative association rules', Cornell University, viewed December 2018, <https://arxiv.org/abs/1806.07084>.
[8] Agarwal, R., Aggarwal, C. and Prasad, V., 2001, 'A tree projection algorithm for generation of frequent item sets', Journal of Parallel Distributed Computing, 61(3), pp.350–371.
[9] Agra, I., Herawan, T., Ghani, N., Ali, A. and Choo, K., 2019, 'A novel association rule mining approach using TID intermediate item set', PLOS ONE Journal, 13(5).
[10] Agrawal, R. and Srikant, R., 1994, 'Fast algorithms for mining association rules in large databases', Proceedings of the 20th VLDB Conference, Santiago, Chile, September 12 - 15, 1994, pp. 487-499.
[11] Agrawal, R., Imielinski, T. and Swami, A., 1993, 'Mining association rules between set of items in large databases', Proceedings of ACM SIGMOD international conference on Management of data, Washington, May 25-28, 1993, pp. 207-216.
[12] Bastide, Y., Taouil, R., Pasquier, N., Stumme, G. and Lakhal, L., 2000, 'Mining frequent patterns with counting inference', Proceedings of ACM SIGKDD international conference on Knowledge Discovery & Data Mining, New York, USA, December 2000, pp.66-75.
[13] Bhargava, R. and Lade, S., 2013,'Effective positive negative association rule mining using improved frequent pattern', International Journal of Modern Engineering Research, 3(2), pp. 1256-1262.
[14] Bhatt, U. and Patel, P., 2015, 'A Novel approach for finding rare items based on multiple minimum support framework', Proceedings of 3rd International conference on recent trends in computing, Elsevier, Vol.57, pp.1088-1095.