Computing and Artificial Intelligence 2025, 3(1), 1819.
https://doi.org/10.59400/cai1819

Article

Multilevel rules mining association for processing big data using genetic algorithm

Gebeyehu Belay Gebremeskel[1,*], Teshale Wubie Yilma[2]

[1] Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar PO BOX 26, Ethiopia
[2] Department of Computer Science, Informatics College, Debre Tabor University, Debra Tabor PO BOX 272, Ethiopia
*** Corresponding author:** Gebeyehu Belay Gebremeskel, gebeyehu.belay@bdu.edu.et

CITATION

Gebremeskel GB, Yilma TW. Multilevel rules mining association for processing big data using genetic algorithm. Computing and Artificial Intelligence. 2025; 3(1): 1819. https://doi.org/10.59400/cai1819

ARTICLE INFO

Received: 12 January 2025
Accepted: 19 February 2025
Available online: 25 February 2025

COPYRIGHT

Copyright © 2025 by author(s). *Computing and Artificial Intelligence* is published by EnPress Publisher, LLC. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: Data mining is a machine learning method and a subset of artificial intelligence that focuses on developing algorithms to enable a computer to learn from data and past experiences within its context. Multilevel association rules mining is a crucial area for discovering interesting relationships between data elements at various levels of abstraction. Many existing algorithms addressing this issue rely on exhaustive search methods such as Apriori and FP-growth. However, these methods incur significant computational costs when applied to big data applications searching for association rules. Therefore, we propose a novel genetic-based method with three key innovations to speed up the search for multilevel association rules and reduce excessive computation. Firstly, we utilize the category tree to describe multilevel application data sets as domain knowledge. Next, we introduce a unique tree-encoding schema based on the category tree to develop the heuristic multilevel association-mining algorithm. Lastly, we present a genetic algorithm based on the tree-encoding schema that greatly decreases the association rule search space. This method is valuable for mining multilevel association rules in big data applications.

Keywords: data mining; market basket data; genetic algorithm; association rules; apriori algorithm; optimization

1. Introduction

In modern business, association rule mining is essential for accessing and managing large-scale datasets in a storage vault. An advanced and dynamic data management system provides an opportunity to expand knowledge of natural business data phenomena. Researchers have conducted association rule mining and classification strategies to locate relevant data in extensive databases. Their research focus and findings rely on vast amounts of data stored in modern databases, which poses significant challenges [1,2]. The databases typically contain many tuples and attributes, requiring more storage space. Therefore, association rule mining with big data is a prominent research method that has revolutionized how organizations view their performance and operation. It allows access to more data than organizations can handle and translates into value [3].

Associations rule mining has become an essential and widely used method for extracting frequent data items, associations, and correlations between data items from datasets [4]. The association system in the market basket is analyzed to identify items frequently purchased together [5]. Based on this data companies can tailor their marketing campaigns to promote additional items and increase revenue. Association and relationship are found in frequent item set discoveries among large data sets [6]. Association rule mining is essential for generating rules with confident estimates of less than one [7]. The main challenge of data mining is to develop a fast and efficient algorithm capable of handling large data sets effectively. Association rule mining is an essential and successful technique to extract valuable information from large databases [8]. It identifies items that have confidence in a minimum support request itemsets that support and uses the frequent items to generate desired rules with confidence. Association rule mining for market basket analysis is used in item set searching and interests [9]. Association rule mining is a two-stage process:

- Find all item sets: Each item must occur at least as frequently as the specified minimum support threshold,
- Generate association rules from the frequent item sets: These rules must meet both the minimum support and minimum confidence criteria; Additional interesting quality measures for rules are determined by the initial step, as illustrated in **Figure 1**.

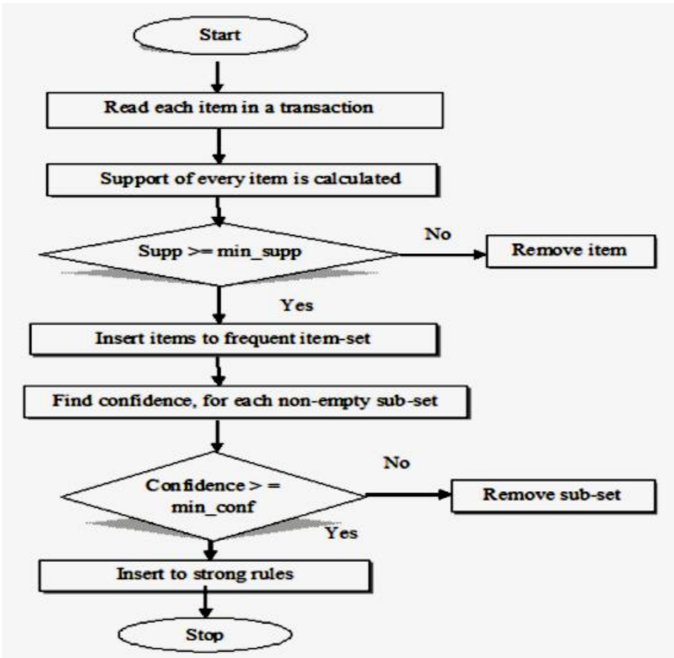

Figure 1. Process in association mining rules.

Various research has been conducted on association rule mining to search for relationships and connections in item data sets in transactional and social databases. For example, in a grocery store, this method is applied to identify items frequently purchased together, which can be used in Market Basket Analysis (MBA) to determine customers' preferred products [10]. Using an MBA is made easier by data mining tools, allowing marketers to identify popular items through an MBA and determine the combined purchase rates of these items [11]. The integrity levels of items refer to how often items are bought together, which can be determined through a database query. For instance, with 100 items, it would take thousands of queries to find the support and confidence association key component descriptions.

- Support measures the percentage of itemsets that fit within a given set of the rule's antecedents and consequents.

- Confidence: It measures how often the consequents are trusted when the antecedents are true.

In some database applications, data is structured hierarchically with different levels of concepts. An optimal hierarchy could reflect the categorization of items sold in an electronics retail store (computers, audio items, etc.). A database may aim to identify ARM among items at the same level or across different levels.

1.1. Association rules for items databases categories

Transactional databases linking association rules to value-based databases are a collection of transaction evidence optimized for computers and e-commerce business processes in large databases. Data mining techniques focus on discovering association rules to identify relationships among items in transaction records. These techniques offer generalized association rule mining with a taxonomy for beneficial common flat association rule mining [12]. The retail item process considers specific standards to categorize item transaction records based on mining-related association rules as new transactions are continuously grouped as the database expands.

1.2. Apriori mining rule

The Apriori Algorithm is a data mining technique used in association analysis to mine association rules [13]. It attempts to search for association rules based on the minimum support and confidence threshold. Support values in rules search for relationships among items greater than or equal to the minimum support threshold and confidence values are greater than or equal to the minimum confidence threshold [14,15]. In items purchasing transactions or purchases in a computer-based analysis, the goal is to find sets of items frequently sold together, known as frequent itemsets. These frequent itemsets are processed to support additional items a current customer might purchase. The Apriori algorithms-based transaction system provides:

- To discover an efficient transaction based on association rules using item support and confidence sets that meet a minimal threshold value measure.
- To build a set of itemsets in various sizes of K items that meet the items support and confidence system using a sequence of K-1 item sizes.

1.3. Association mining rules for market basket analysis standards

Association rule mining for market basket analysis involves examining collections of items within a single transaction. The quantity of each item purchased in a transaction is less important than the variety of items bought. The focus is on analyzing the combined transactions of multiple customers over an extended period. A market basket refers to a set of items purchased together by a customer during a single visit to a store.

1.3.1. Enumeration data

Association rule mining-based analysis in market basket standard uses enumerated data to provide a wide range of analytical information about customers. The data is related to customers and business assessments used in the planning process such as health, education, transportation, and stores. These public businesses establish new businesses or financial industries. Data mining techniques are applied to enumerate data to manage transaction data providing significant support for good public item set similarity searching rules and effectively facilitating the functioning of the business process.

1.3.2. Agribusiness

In precision agriculture, agribusiness datasets are crucial for improving resource allocation, community planning, and business transactions. These datasets rely on efficient and timely research methods. Association rule mining is utilized to explore relationships between itemsets in a large database to understand customer purchasing behavior and increase sales [16]. This method can assist in selecting the most appropriate products for individual users based on current marketing conditions and geographical location ultimately maximizing business transaction performance.

1.4. The problem with the statement

The standard rule mining procedures have numerous constraints when inferring relationships in large-scale databases. In a traditional mining approach accuracy is a significant measuring concern [17]. Additionally, the time required to find data relationships increases when dealing with large-scale datasets. Association rule mining with big data is a data mining technique used to uncover hidden data in large-scale datasets. Association rules effectively reveal interesting relationships in large itemsets transactions that collocate within large item datasets. The data that occurs through association rules mining illustrates similarities in large transactions [3]. This study focuses on the uncertainty of Apriori algorithms, which require multiple scans of databases to compute time for item transactions [5].

Various issues with Apriori algorithms exist in current systems focusing on item transactions impacting overall transaction effectiveness. These challenges can be addressed differently, including (i) The reliability and consistency of current Market Basket Analysis methods are insignificant; (ii) Most existing methodologies are not adaptive, leading to long processing times to mine appropriate items support and confidence; (iii) Continuity and repetition in search results for Irrelevant items are common. Therefore, Association rule mining presents a novel Market Basket Analysis approach to overcome these obstacles. Additionally, association rule mining techniques for large itemsets can increase sales by mining association rules at multiple levels and help discover item relationships among transactions. To address these gaps and problems, the following research questions are formulated.

- How can association rules be efficiently mined from large databases, and what type of data is most suitable for association rules mining?
- How can genetic algorithms be used as a problem-solving technique in data mining to search for item relationships in large databases and association rule mining processes?
- What is the best method for measuring associations between attributes of interest?

Multi-level association rules aim to explore association rule-based mining algorithms for Market Basket Analysis [11]. The research specifically addresses data handling to apply association rules in large-scale data and build a proficient association rule mining strategy to provide better proposal results that fulfill the client's needs [18]. It enhances the association rule mining application strategy of item databases by selecting the most similar items for efficient transactions in proper business conditions and their association [19]. These methodological rule-mining techniques aim to identify significant relationships within a large dataset of data items. Data mining techniques are powerful and advanced technologies that focus on extracting hidden predictive data [20]. This system has an impact on assisting various associations in focusing on large-scale data [1]. The proposed technique can easily predict future trends and practices of new advancements, helping organizations make proactive and data-driven decisions [21]. The contribution of the paper is summarized as follows.

- Adequately handling substantial databases to find association rules with high accuracy,
- Evaluating interesting patterns representing data based on quality measures,
- Enhancing evaluation and interpretation integrity of rules mining and reliability to define goals of item transactions from the initial phase to the end,
- The perception and data representation for big data become dynamic, implying potential improvements to the system and measuring the impacts.

This paper is structured as follows: Section 1 provides an introduction outlining the paper's objective, research focus, and anticipated contributions. Section 2 presents a detailed review of related work. Section 3 elucidates the research methodology. Section 4 presents the experimental setup, results, and discussion. Section 5 concludes the paper, followed by a list of references.

2. Related works

Mining Association Rules is the most essential field of application for data mining [9]. Association rule mining is a specific application of MBA [11], where retail transaction data is analyzed to discover items likely to be purchased together. Research has focused on a grocery store scenario database for purchased items in single-time customer transactions [19,22], developing a Predictive Association Rules (PAR) technique that combines the benefits of traditional rule-based classification. Rule mining classification is an association rule mining strategy that determines how to build an ARM-based classifier using Classification Association Rules (CARs) [23].

Big data combined with association rule mining can identify the most influential factors beyond what conventional and official statistics metrics can provide. These complements can address the gaps and deficiencies in official data and measurement methods [10]. This process involves the classification and performance measurement of large data sets. Many researchers have explored how the characteristics of market items associated with selected market baskets can generate advanced mining Interesting Rules through association and classification algorithms [24]. However, a drawback is that some aspects of the generated rules may not be relevant to certain users. Therefore, certainty support, lift confidence, and gain have been proposed to identify the best or most interesting rule-mining principles [25]. Spatial interpolation is recommended to create a continuous LMAS that can be used to analyze hotspots displaying strong association patterns [26].

The concept of cross-section and continuous itemsets is utilized, in the PRuleed Concept Lattice (PCL) technique to identify regular itemsets in a predefined database, and effectively compress the size of itemsets [27]. Studies on mining transaction association rules reveal items related to the same transaction [28]. A key aspect is leveraging the confidence required for the resulting itemset, ensuring only necessary itemsets are generated. The Apriori approach [29] in multilevel association rules mining involves including all ancestors of frequent items in the transaction database. This accumulates, exhaustively identifying frequent items at each concept level in the ML-T2L1 algorithm [30,31] and the Level-Crossing algorithm [32].

To enhance computing efficiency, Preeti and Nidhi, [32] optimized the functions of each item and computed the minimum support. GA-based methods can efficiently scan a large number of association rule candidates. According to [33], GA-based methods can uncover high-level prediction rules combined with multilevel association rule mining. This approach is proposed as an attempt to efficiently discover multilevel association rules in market basket data [34].

3. Methodology

Association rules mining involves multi-level itemsets in a catalog tree that outlines transactions based on item relationships [35]. Multilevel rule mining is an interesting issue in the item domain knowledge represented as parent-child relationships. These relationships flow from the supper class to subsequent child nodes or an edge from one node to another annotated as î an ancestor of one and as a descendant of î, the other creating a path between them, with their leaf nodes included in the datasets. For example, in a supermarket domain, a classification tree shown in **Figure 2** represents a transactions dataset where each node consists of a set of items. Each transaction is linked to an identifier, and items are expected to be present based on transaction supports within the transaction itself or as an ancestor of some items [36].

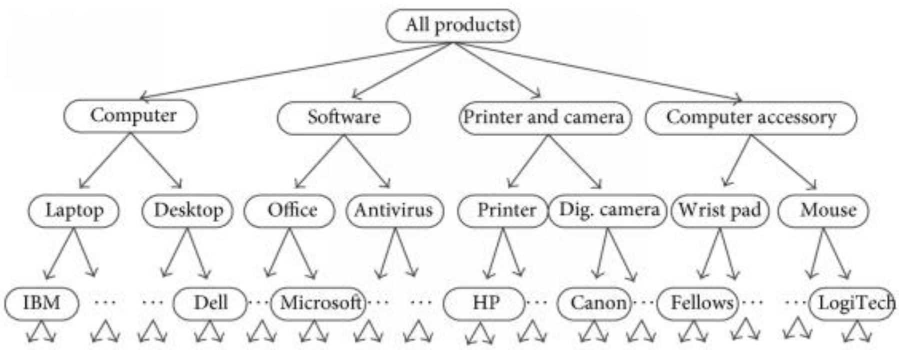

Figure 2. Supermarket itemsets catalog tree.

Figure 2 shows the rule mining application for supermarket product transactions with the support of a classifier. The term support represents the percentage of transactions in which items are sold together. Additionally, the concept of item confidence in a transaction set is where confidence is the percentage of item transactions supported to all business flows. This can be viewed as the conditional probability of mining multilevel association rules refined through the process of all itemsets. The correlation of the transaction database involves GA techniques such as item category roll-up, mutation, crossover, and selection operators [37]. A transaction supports items, if it supports every item in the transactions. Association rule mining evaluates the multilevel support and confidence abstraction by considering the interests of the items.

- Support Confidence: These rules focus on items with high support and confidence in transactions indicating more interesting rules with high confidence to discover in the itemset database.
- Interest: This rule is suitable for preferring the catalog tree, and association mining-based item relation search is essential to improve item support.

Multilevel association rule mining at a primitive dataset level makes it challenging to gain correlated item-sets transactions. In a large-scale database ltemsets records present in the big data context, which increases the number of items in the catalog and transactions. In traditional methods computation and memory consumption expand exponentially [38]. This process is important to note the FP-tree algorithm to improve the association rules mining efficiency in multilevel association rules. It is specific but also cross-level association rules to enhance item selection support and confidence. Therefore, big data analysis is a novel and heuristic method for multilevel association rules mining [39].

3.1. Proposed genetic algorithm-based association generation method

3.1.1. General schema of genetic algorithm

A genetic algorithm (GA) based heuristic and an adaptive search strategy are used to search and optimize procedures, inspired by principles of natural selection and genetics [40]. GA is a part of evolutionary computing and it is utilized to solve optimization problems that mimic the survival of the fittest among entities over successive generations [41]. According to [6], the key reason for using gas in association rules mining is the comprehensive search and handle attribute interface better than greedy rule induction algorithms [42]. In multilevel association rule mining Genetic algorithms can limit the association rule search space and achieve optimal results during association rule discovery [43]. Typically, for association rule mining all candidate item sets need to be generated and checked against the entire dataset using GA processes to solve problems following these steps [44].

- Initialized population: The population chromosomes represent a set of candidates for searching for a solution to the problem.
- Evaluate each individual in the population: Calculate the individual candidate in the population for the fittest one.
- Repeat
- Select parents: This strategy is to select parents on their fitness.
- Recombination of parent pairs: Recombination is achieved through a genetic operator called (crossover), which exchanges subsequences of mating chromosomes.
- Mutate the resulting offspring: Mutation is another genetic operator that randomly flips some bits in chromosomes based on a certain probability.
- Evaluate offspring: Calculate the fitness of the new generation of chromosomes.

- Insert offspring in the population: The population is reconstructed with new chromosomes.
- Until (stopping criteria): The algorithm stops based on certain constraints.
- Extract solution from population: After the algorithm stops desired solutions are extracted based on certain conditions.

3.1.2. GenRul algorithm

This algorithm is designed to evolve itemsets for association rules [45]. It starts by generating random itemsets from the database, which are then developed to identify items that meet the minimum support threshold. It is important to note that this algorithm specifically deals with categorical attributes. The General process can be completed in multiple steps.

Encoding the problem

As we know the type of encoding used to represent the problem is related to the application of the problem [46]. The challenge involves itemsets that consist of categorical attributes. The suggested encoding is to represent the chromosome that consists of these attributes, with the chromosome divided into multiple areas (similar to bit string). The value associated with each area is constrained to the domain of the attributes [47]. These attributes are called cardinality attributes, and the size of the domain of each area in a chromosome varies. Let the initial population consist of four chromosomes that can be generated. The fitness function is represented by the support value of the object that contains these attributes in the dataset, shown in **Table 1**.

Table 1. Sample chromosomes.

	Outlook	**windy**	**class**
Chromosome1	rain	false	Play
Chromosome2	rain	true	Play
Chromosome3	sunny	true	Play
Chromosome4	overcast	false	Don't play

The desired itemsets involve three attributes (outlook, windy, and class). This means that the desired chromosome contains three partitions representing loci. The initial population consists of four chromosomes.

Input parameter

The general rule requires certain parameters that constraint the evolving process before it begins, such as:

- The size of a chromosome refers to the number of divisions the chromosome must have.
- The minimum support for the itemsets.
- The stopping criteria to end the evolving process.
- The mutation rate.

Fitness function

General aims to find itemsets (combinations of attributes) with high frequencies (support). Therefore, the general algorithm calculates individual fitness based on possible combinations of attributes' states represented in the data set as per algorithm 1.

Algorithm 1 Input data sets attributes combination

```
-Input the size of the chromosome (number of attributes that evolve).
-Build the initial population of chromosomes.
-Repeat until (stopping condition).
{
-For each chromosome (itemsets), compute its fitness (support value for the itemsets).
Extract the itemsets that satisfy the minimum support to constitute the association rules.
Choose randomly the number of chromosomes (itemsets) that satisfy the minimum support to be a parent for
reproducing offspring.
Select the divisions of the pairs of chromosomes to be exchanged. Mutate-produced offspring if the time of mutation
operation is less
than mutation rate.
Evaluate the resulting offspring.
Insert offspring in the population.
}
```

3.1.3. Genetic algorithm-based approach for generating associations

GA operators such as *selection*, *crossover,* and *mutation* are applied on an initially random population to compute a whole generation of new strings to generate solutions for successive finesses [48]. Success is the probability of an individual reproducing, and the goodness of the representation of proportional solutions through the improved quality of the solutions in successive generations [49]. The process is terminated when an acceptable or optimum solution utilizes GA operators in appropriate processes for problems from the optimal and computable criterion of functions [50,51]. The genetic operator's representation is described as follows.

Selection: This deals with selecting the fittest chromosome, using probabilistic Fitness to measure how well a chromosome solves the issue.

Crossover: The crossover operation is a random selection that occurs along with the length of the chromosomes, swapping all the genes after that point.

Mutation: The process is gaining new solutions of mutations to address stochasticity in the search, and the chance that a bit within an inverted chromosome (0 becomes 1, 1 becomes 0).

The GA-based operators perform based on frequent itemsets searching to transform the population by executing the various and repetitive steps (**Figure 3**):

- *Fitness Evaluation*: Fitness is a search finding (i.e., an objective function) for each individual.
- *Selection*: This genetic operation works by individual selecting from the current population to be involved in recombination.
- *Recombination*: It is new offspring that are created from parents using genetic operators like crossover and mutation.
- *Replacemen*t: This is searching for offspring with individuals (usually their parents).

A single cycle transforms a population in a replaced process with offspring. This transforming cycle of the entire population is called the generation gap (between 0 and 1) algorithm 2. The GA pseudo code also describes input, and output components using the following genetic algorithm technique.

Algorithm 2 Genetic algorithm pseudocode

```
Input: Dataset (DB0), association rules mining (ARMs), candidate transactions (CTs), non-sensitive association rules
(NSARs), sensitive association rule (SAR), d victim Item (VI), MS, MC
Output: Global best individual
Start
1. initial population = Randomly initialize population
2. FitnessValues = objective function (initial population, DB0, CT, VI, AR, NSAR, MS, MC, d)
3. while (! stop condition)
4. bestFitIndividuals = selectElite (FitnessValues)
5. new individuals = crossover and mutation (bestFitIndividuals)
6. FitnessValues = objective function (new individuals, DB0, CT, VI, AR, NSAR, MS, MC, d)
7. end while
8. globalBestIndividual = individual with min(FitnessValues)
9. FitnessValues = objective function (Population, DB0, CT, VI, AR, NSAR, MS, MC, d)
10. For each solution from the Population
11. NewDB0 = DeleteVictim (DB0, VI, solution)
12. LostRules = FindLostRules (NewDB0, DB0)
13. LNSAR = FindLostNonSensitive (LostRules, NSAR)
14. FitnessValues. Add(LNSAR/NSAR)
15. end
End
```

The Adopted Genetic Algorithm pseudocode.

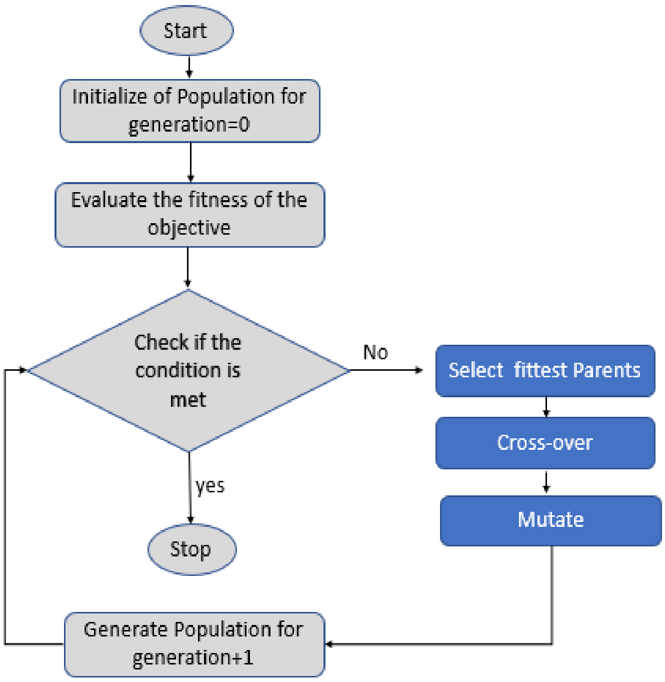

Figure 3. Proposed genetic algorithm workflow.

Initialize the population for Generation 0: GA operates by creating an initial population through a random bias and possible candidate solutions for a problem. This solution is evaluated in the generation based on the fitness function, and the termination condition as follows:

- Generate predefined generation using GA.
- Predefined target value from the fitness function.

The fittest population members are identified, ranked, and selected as parents for the next generation when the termination condition is not met, generating the population for the generation: The genotype of the next generation is also created through genetic operations such as crossover and mutation as shown in **Figure 4**.

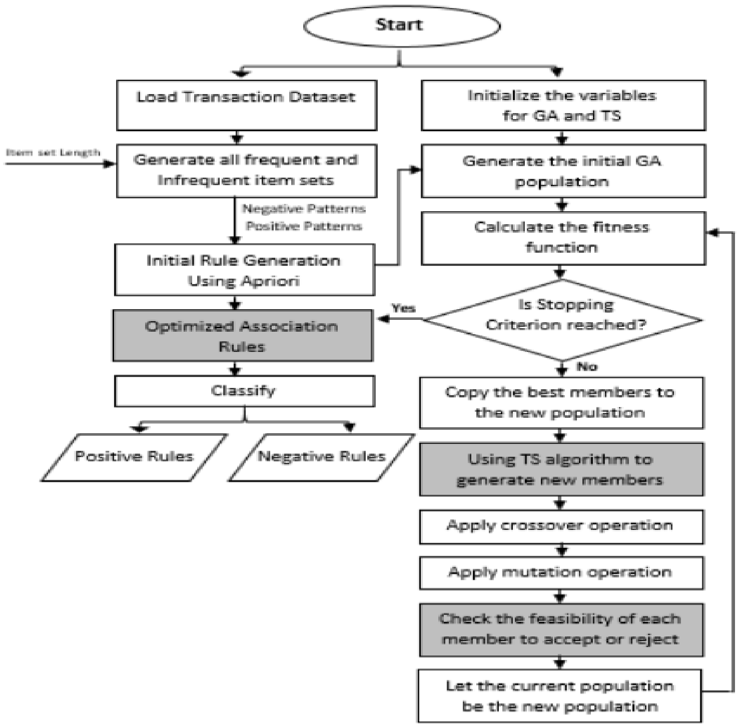

Figure 4. Architecture of genetic algorithm-based approach for association generation.

Genetic algorithms are optimization techniques inspired by biological evolution. They employ selection, crossover, and mutation iterative solutions [52,53] . Unlike traditional methods, genetic algorithms explore the search space efficiently and adaptively, making them well-suited for complex problems. In multilevel association rule mining, genetic algorithms offer a more efficient approach [54]. Instead of exhaustively testing all potential item combinations, they can intelligently generate and evaluate candidates based on their evolutionary fitness. This reduces computational overhead and improves scalability.

3.1.4. Encoding scheme

A crucial step in using genetic algorithms for multilevel association rule mining is representing association rule candidates in a suitable encoding. Traditional encodings often fall short of this task. We propose a novel category tree-based encoding scheme. In this approach, each valid multilevel association rule is represented as a subtree within the category tree [55]. Each leaf node is assigned a binary value to signify whether it's part of the rule's antecedent or consequent. This encoding allows for efficient representation and manipulation of association rule candidates within the genetic algorithm framework.

In the GA-based encoding subtree, leaf nodes represent commodities and are assigned a 0, 1, or −1. Commodities with a value of 0 are included in the rule's antecedent, while those with a value of 1 are part of the consequent. Commodities assigned -1 are excluded from the rule. To initialize the first generation of association rules, we randomly prune the category tree and assign values to the remaining leaf nodes, creating subtrees that represent potential association rules.

3.1.5. Genetic operators

Starting with the provided category tree, we randomly prune it to create subtrees that serve as initial association rule candidates. Each subtree is then assigned random values of −1, 0, or 1 to its leaf nodes, ensuring that both 0 and 1 are present. This process generates a suitable initial population. The selection operator determines which individuals will reproduce. We employ the roulette wheel selection method, where individuals with higher fitness are more likely to be chosen [56]. This approach favors promising association rules for the next generation.

The mutation operator is essential for preserving genetic diversity within the population. To achieve this, we implement three distinct mutation types: randomly selecting a leaf node and assigning it a different attribute value, randomly choosing a non-root node and pruning its entire subtree, and randomly selecting a non-root node and attaching a randomly generated subtree. These mutations help explore the search space and prevent premature convergence [57].

3.1.6. Fitness function

The fitness function is a fundamental component of our genetic algorithm. It assesses the quality of each offspring generated and serves as a stopping criterion when a sufficient number of high-quality association rules have been produced. The fitness of an association rule is determined by two key factors: support and confidence. To balance the importance of factors and to account for the varying complexity of association rules at different levels of the category tree, we introduce a novel approach: level-specific minimum support thresholds [58]. This means the deeper levels of the association rule tree have lower minimum support requirements than shallower levels. This strategy allows for a more nuanced evaluation of association rules, as it recognizes that more complex relationships may have lower support but still be valuable.

4. Experiment and result discussion

Association rule mining (ARM) has become a valuable tool for marketers analyzing customer behavior. By examining market baskets, ARM identifies frequently purchased item combinations. Multiple datasets are used to uncover hidden relationships between products. This analysis helps marketers understand customer preferences and tailor marketing strategies accordingly. Various ARM algorithms are compared to determine their effectiveness. This comparative analysis provides insights into different approaches and their suitability for specific applications as shown in **Figures 5–7**.

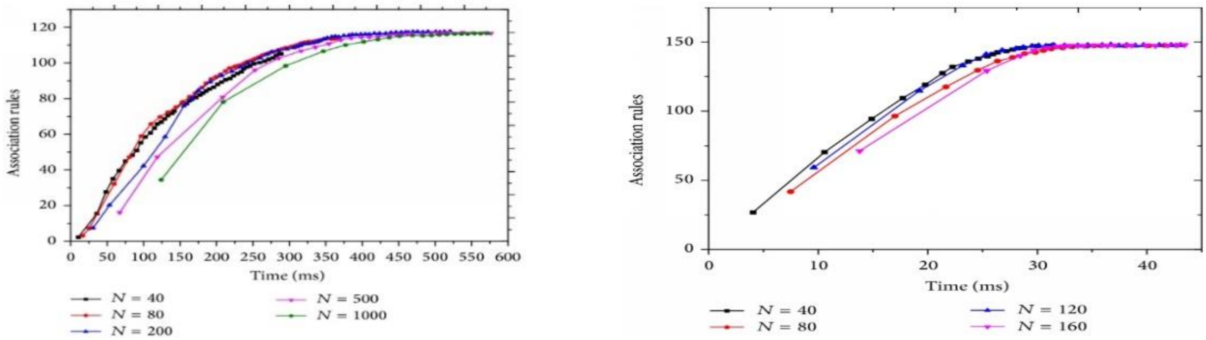

Figure 5. Dataset1&2 observation: With a limited population and a limited time, most valid association rules could be mined in dataset 1.

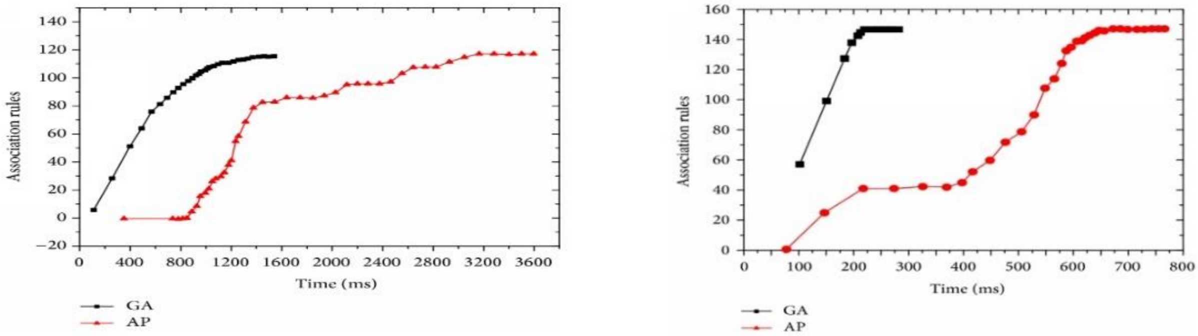

Figure 6. Number of association rules mined from 4000 transactions in dataset 1 & 2.

Observation: The GA-based approach consistently outperforms the Apriori algorithm's speed and efficiency. While Apriori may potentially discover more rules given unlimited time, the GA-based approach provides a significant advantage in practical scenarios with time constraints.

Table 2. Average fitnesses of association rules mined from 500 and 1000 transactions in two datasets.

A result from datasets 1								
Number of transactions	**Type**	**Times**						
		62.5	**125**	**250**	**400**	**1000**	**2000**	**2500**
1000	GA	32.8	67.7	94.4	97	97	97	97
	AP	0	0	0	46.4	130	151	151
2000	GA	7	30.4	66.9	92.3	116.6	116.6	116.6
	AP	0	0	0	0	35	97	118
A result from dataset 2								
Number of transactions	**Type**	**Times**						
		62.5	**125**	**250**	**400**	**1000**	**2000**	**2500**
500	GA	57.6	97.2	150.9	151	151	151	151
	AP	0	0	0	46.4	130	151	151
1000	GA	0	57.3	130.9	147	147	147	147
	AP	0	0	0	39	74	141	147

Table 3. Average fitnesses of association rules mined from 1000 and 2000 transactions in two datasets.

A result in datasets 1									
Number of transactions	**Type**	**Time(s)**							
		20	**62.5**	**125**	**250**	**400**	**1000**	**2000**	**250**
1000	GA	0.14	0.17	0.17	0.16	0.16	20	62.5	125
	AP	0	0	0	0.13	0.15	0.15	0.16	0.16
2000	GA	0	0.21	0.24	0.23	0.22	0.20	0.20	0.20
	AP	0	0	0	0	0	0.16	0.20	0.20
A result in datasets 2									
Number of transactions	**Type**	**Time(s)**							
		15	**20**	**40**	**100**	**200**	**400**	**600**	**1000**
500	GA	0.35	0.36	0.35	0.35	0.35	0.35	0.35	0.35
	AP	0	0	0	0.29	0.36	0.35	0.35	0.35
1000	GA	0	0.36	0.37	0.35	0.35	0.35	0.35	0.35
	AP	0	0	0	0.26	0.28	0.36	0.35	0.35

Tables 2 and **3**: Observation: When analyzing 1000 and 2000 transaction records, the GA-based approach consistently outperformed the Apriori algorithm in terms of efficiency and accuracy in discovering valid association rules. These results, as illustrated in **Figure 7**, demonstrate the superiority of the GA-based approach.

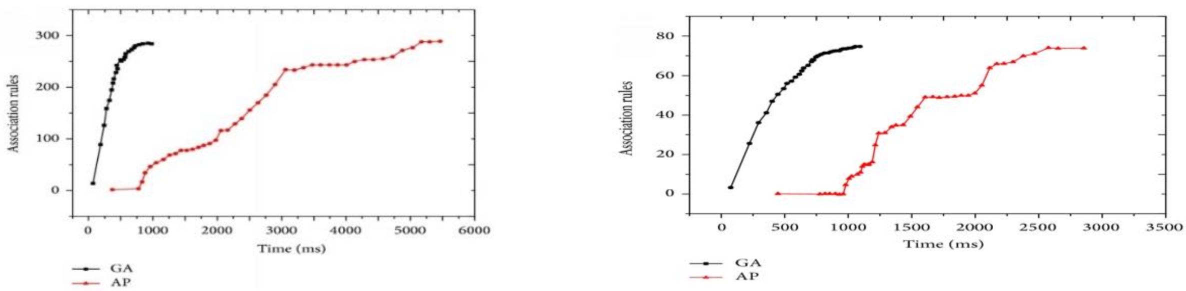

Figure 7. GA-based and apriori algorithms with different min_sup and min_conf in datasets 1 & 2.

Observation: Adjusting the minimum support and confidence thresholds significantly impacted the output of both the GA-based and the Apriori algorithms.

- Dataset 1: Reducing min_sup to 0.005 and min_conf to 0.25 resulted in the GA-based algorithm discovering more association rules in a shorter time to Apriori.
- Dataset 2: Increasing min_sup to 0.02 and min_conf to 0.75 yielded similar results, with the GA-based algorithm outperforming Apriori.

These findings consistently demonstrate the superior efficiency and effectiveness of the GA-based approach in discovering association rules under various threshold settings as shown in **Figure 8**.

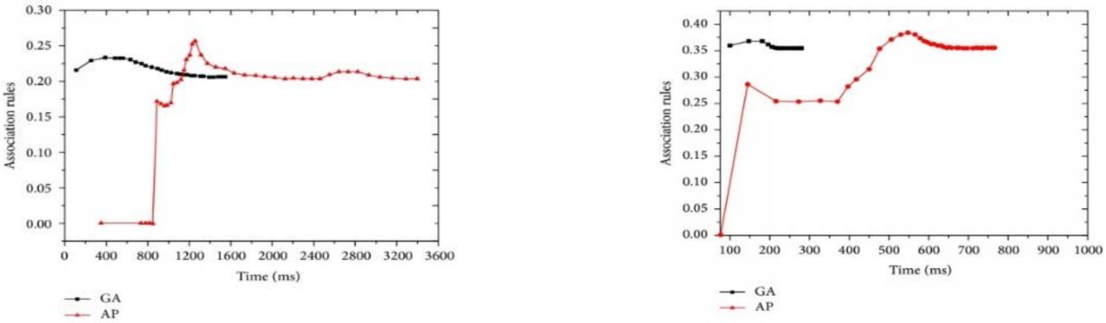

Figure 8. Average fitness of association rules mined from 4000 transactions in dataset 1. & 2.

Observation: To evaluate the value of the mined association rules, we applied the same fitness function to both the GA-based and the Apriori algorithms. Experiments were conducted on datasets with varying transaction sizes. In previous findings, the GA-based approach consistently identified association rules in a shorter time compared to the Apriori algorithm across all datasets. This demonstrates the superior efficiency and quality of the association rules discovered by the GA-based approach.

5. Conclusion

We have presented a novel genetic algorithm-based method for mining multilevel association rules in large data sets. This paper presents a novel genetic algorithm (GA)--based approach for efficiently mining multilevel association rules from large datasets. By leveraging a catalog tree to represent domain knowledge, we have developed a specialized subtree-based encoding scheme that enables the GA to explore the search space. We have also tailored the initialization, crossover, and mutation operators to suit this tree-based representation. Multilevel association rule mining often suffers from high computational costs due to exhaustive database scans. Our GA-based approach addresses these challenges by employing a dynamic fitness function on multilevel support and confidence thresholds. This adaptability allows our algorithm to converge efficiently, even in large datasets. While our method demonstrates significant promise, there are areas for future exploration. Our approach is particularly effective in domains where items can be organized. However, it may not be as suitable for unstructured item sets. Additionally, exploring distributed and parallel computing implementations could further optimize the performance of our GA-based algorithm. Future research could focus on adapting various algorithms to generate association rules that enhance existing recommendation systems. By incorporating these advancements, recommendation systems can become more practical and efficient, providing personalized experiences to users.

Author contributions: Conceptualization, GBG and TWY; methodology, GBG and TWY; software, GBG and TWY; validation, GBG; formal analysis, GBG; investigation, GBG and TWY; resources, GBG and TWY; data curation, TWY; writing—original draft preparation, GBG and TWY; writing—review and editing, GBG; visualization, GBG; supervision, GBG; project administration, GBG. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: We thank the anonymous reviewers for their valuable comments and constructive suggestions. Any public, commercial, or not-for-profit organizations did not fund this research.

Conflicts of interest: The authors declare no conflict of interest.

References

1. Zakur Y, Flaih L. Apriori Algorithm and Hybrid Apriori Algorithm in the Data Mining: A Comprehensive Review. E3S Web of Conferences. 2023; 448: 02021. doi: 10.1051/e3sconf/202344802021
2. Setiabudi DH, Budhi GS, Purnama IWJ, et al. Data mining market basket analysis’ using hybrid-dimension association rules, case study in Minimarket X. In: Proceedings of the 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering; August 2011.
3. Sigala M, Rahimi R, Thelwall M, et al. Big Data and Innovation in Tourism, Travel, and Hospitality. Springer Singapore; 2019.
4. Sharma A, Ganpati A. Association Rule Mining Algorithms: A Comparative Review. International Research Journal of Engineering and Technology (IRJET). 2021; 8(11): 848-853.
5. Wahidi N, Ismailova R. Association rule mining algorithm implementation for e-commerce in the retail sector. Journal of Applied Research in Technology & Engineering. 2024; 5(2): 63-68. doi: 10.4995/jarte.2024.20753
6. Grami M, Gheibi R, Rahimi F. A novel association rule mining using genetic algorithm. In: Proceedings of the 2016 Eighth International Conference on Information and Knowledge Technology (IKT); September 2016.
7. Moslehi F, Haeri A. A genetic algorithm-based framework for mining quantitative association rules without specifying minimum support and confidence. Scientia Iranica D. 2020; 27(3): 1316-1332.
8. Zhang M, Fan J, Sharma A, et al. Data mining applications in university information management system development. Journal of Intelligent Systems. 2022; 31(1): 207-220. doi: 10.1515/jisys-2022-0006
9. Selvi RS, Valarmathi ML. Optimal Feature Selection for Big Data Classification: Firefly with Lion-Assisted Model. Big Data. 2020; 8(2): 125-146. doi: 10.1089/big.2019.0022
10. Saxena A, Rajpoot V. A Comparative Analysis of Association Rule Mining Algorithms. IOP Conference Series: Materials Science and Engineering. 2021; 1099(1): 012032. doi: 10.1088/1757-899x/1099/1/012032
11. Karapiperis C, Chasapi A, Angelis L, et al. The Coming of Age for Big Data in Systems Radiobiology, an Engineering Perspective. Big Data. 2021; 9(1): 63-71. doi: 10.1089/big.2019.0144
12. Hashad AA, Khai Wah K, et al. Exploratory analysis with association rule mining algorithms in the retail industry. Malaysian journal of computing (MJOC). 2024; 9(1): 1746-1758. doi: 10.24191/mjoc.v9i1.21433
13. Shrivastava A, Jain RC. Performance Analysis of Modified Algorithm for Finding Multilevel Association Rules. Computer Science & Engineering: An International Journal. 2013; 3(4): 1-9. doi: 10.5121/cseij.2013.3401
14. Zhen H, Chiou BC, Tsou YT, et al. Association Rule Mining with Differential Privacy. In: Proceedings of the 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W); June 2020.
15. Rawat SS, Rajamani L. Probability apriori based approach to mine rare association rules. In: Proceedings of the 2011 3rd Conference on Data Mining and Optimization (DMO); June 2011.
16. Bagui S, Stanley P. Mining frequent itemsets from streaming transaction data using genetic algorithms. Journal of Big Data. 2020; 7(1). doi: 10.1186/s40537-020-00330-9
17. Dolores M, Fernandez-Basso C, Gómez-Romero J, et al. A big data association rule mining based approach for energy building behaviour analysis in an IoT environment. Scientific Reports. 2023; 13(1). doi: 10.1038/s41598-023-47056-1
18. Kelotra A, Pandey P. Stock Market Prediction Using Optimized Deep-ConvLSTM Model. Big Data. 2020; 8(1): 5-24. doi: 10.1089/big.2018.0143
19. Zhao Z, Jian Z, Gaba GS, et al. An improved association rule mining algorithm for large data. Journal of Intelligent Systems. 2021; 30(1): 750-762. doi: 10.1515/jisys-2020-0121
20. Kaur G, Singh L. Data Mining: An Overview. International Journal of Computer Science and Telecommunications. 2011; 2(2): 336-339.
21. Wu H. Data Association Rules Mining Method Based on Improved Apriori Algorithm. In: Proceedings of the 2020 the 4th International Conference on Big Data Research (ICBDR’20); 27 November 2020.

22. Bao F, Mao L, Zhu Y, et al. An Improved Evaluation Methodology for Mining Association Rules. Axioms. 2021; 11(1): 17. doi: 10.3390/axioms11010017
23. Wu X, Wen Q, Zhu J. Association rule mining with a special rule coding and dynamic genetic algorithm for air quality impact factors in Beijing, China. PLOS ONE. 2024; 19(3): e0299865. doi: 10.1371/journal.pone.0299865
24. Aggarwal H, Kumar V, Arora HD. Data mining algorithm based on Renyi fuzzy association rule: an application for selecting a suitable course. Research in Statistics. 2023; 1(1). doi: 10.1080/27684520.2023.2271902
25. Darwish SM, Essa RM, Osman MA, et al. Privacy-Preserving Data Mining Framework for Negative Association Rules: An Application to Healthcare Informatics. IEEE Access. 2022; 10: 76268-76280. doi: 10.1109/access.2022.3192447
26. Malamsha GC, Nyambo DG. Multi-level Association Rule Mining for the Discovery of Strong Underrepresented Patterns. Engineering, Technology & Applied Science Research. 2023; 13(2): 10377-10383. doi: 10.48084/etasr.5683
27. Yildirim Taşer P, Birant Ku, Birant D. Multitask-based association rule mining. Turkish journal of electrical engineering & computer sciences. 2020; 28(2): 933-955. doi: 10.3906/elk-1905-88
28. Raorane A, Kulkarni RV. Data Mining Techniques: A Source for Consumer Behavior Analysis. International Journal of Database Management Systems. 2011; 3(3): 45-56. doi: 10.5121/ijdms.2011.3304
29. Mudumba B, Kabir MF. Mine-first association rule mining: An integration of independent frequent patterns in distributed environments. Decision Analytics Journal. 2024; 10: 100434. doi: 10.1016/j.dajour.2024.100434
30. Schlegel B, Gemulla R, Lehner W. Memory-efficient frequent-itemset mining. In: Proceedings of the 14th International Conference on Extending Database Technology; 21 March 2011.
31. Kishor P, Porika S. A Novel Association Rule Mining Model for Generating Positive and Negative Association Rules with Hybridized Meta-Heuristic Development. International Journal of Intelligent Engineering and Systems. 2023; 16(2): 125-141.
32. Zheng L. Research on E-Commerce Potential Client Mining Applied to Apriori Association Rule Algorithm. In: Proceedings of the 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS); January 2020.
33. Yusof R, Manaf M, Mohd ZNA, et al. The Complexity of Extracting Knowledge in Big Data. Journal of Mathematical and Computational Science. 2016; 1(1): 12-18.
34. Xu Y, Zeng M, Liu Q, et al. A Genetic Algorithm Based Multilevel Association Rules Mining for Big Datasets. Ding S, ed. Mathematical Problems in Engineering. 2014; 2014(1). doi: 10.1155/2014/867149
35. Hibbi FZ, Abdoun O, Haimoudi EK. Exploration of Analytical Mechanisms in the Feedback model. Procedia Computer Science. 2019; 148: 201-207. doi: 10.1016/j.procs.2019.01.037
36. Jain A. Association Rule Mining in Transactional Data: Challenges and Opportunities. International Journal of Mechanical Engineering. 2021; 6(3): 4548-4558.
37. Xie L, Chen Z, Yu S. Deep Convolutional Transformer Network for Stock Movement Prediction. Electronics. 2024; 13(21): 4225. doi: 10.3390/electronics13214225
38. Yan X, Zhang C, Zhang S. Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Systems with Applications. 2009; 36(2): 3066-3076. doi: 10.1016/j.eswa.2008.01.028,
39. Nurmayanti WP, Sastriana HM, et al. Market Basket Analysis with Apriori Algorithm and Frequent Pattern Growth (Fp-Growth) on Outdoor Product Sales Data, 2021. International Journal of Educational Research and Social Science. 2021; 2(1): 132-139.
40. Zhao Z, Jian Z, Gaba GS, et al. An improved association rule mining algorithm for large data, De Gruyter. Journal of Intelligent Systems. 2021; 30: 750–762.
41. Aqra I, Abdul Ghani N, Maple C, et al. Incremental Algorithm for Association Rule Mining under Dynamic Threshold. Applied Sciences. 2019; 9(24): 5398. doi: 10.3390/app9245398
42. Susan S, Bhutani A. Data Mining with Association Rules for Scheduling Open Elective Courses Using Optimization Algorithms. Advances in Intelligent Systems and Computing. 2020; 941: 770–778.
43. Darwish SM, Amer AA, Taktak SG. A Novel Approach for Discovery Quantitative Fuzzy Multi-Level Association Rules Mining Using Genetic Algorithm. International Journal of Advanced Research in Artificial Intelligence. 2016; 5(6): 35–44.
44. Albattah W, Khan RU, Alsharekh MF, Khasawneh SF. Feature Selection Techniques for Big Data Analytics. Electronics. 2022; 11(19): 3177. doi: 10.3390/electronics11193177
45. Xie HY. Research and Case Analysis of Apriori Algorithm Based on Mining Frequent Item-Sets. Open Journal of Social Sciences. 2021; 9: 458-468. doi: 10.4236/jss.2021.94034

46. Jagli DRD, Sangeeta O. Association Rule Mining: Improved Tree-Based and Graph-Based Approach for Mining Frequent Item Sets. SSRN; 2024.
47. Osuntokun OD, Adeyemo AB, Makolo AU. A Review on Biclustering Algorithms for Data Mining Analysis of Gene Expression Data. Uijslictr. 2020; 5(1): 64-76.
48. Aljehani SS, Alotaibi YA. Preserving Privacy in Association Rule Mining Using Metaheuristic-Based Algorithms: A Systematic Literature Review. IEEE Access. 2024; 12: 21217-21236. doi: 10.1109/access.2024.3362907
49. Hasan AA, Bari QH, Lorber P, et al. An Association Rule Mining approach to explore the dynamics in the plastic recycling business. Cleaner Waste Systems. 2024; 9: 100186. doi: 10.1016/j.clwas.2024.100186
50. Wahidi N, Ismailova R. Association rule mining algorithm implementation for e-commerce in the retail sector. J Appl Res Eng Technol & Engineering. 2024; 5(2): 1-6.
51. Guo F. Data mining techniques for customer relationship management. In: Proceedings of the International Conference on Communication and Electronic Information Engineering (CEIE 2016); 2017.
52. Ramya HR, Pradeep M. A Novel Method for Association Rule Mining using Genetic Algorithm. International Journal for Scientific Research & Development. 2015; 3(4): 533-537.
53. Sharma A, Tiwari N. A Survey of Association Rule Mining Using Genetic Algorithm. International Journal of Computer Applications & Information Technology. 2012; 1(2): 5-11.
54. Sarma R, Sarma PKD, Mazumdar N. Study of Algorithms for Mining Fuzzy Association Rules and Applications. Journal of Electrical Systems. 2024; 20(10s): 5812-5834. doi: 10.52783/jes.6478
55. Machfudiyanto RA, Chen JH, Latief Y, et al. Applying Association Rule Mining to Explore Unsafe Behaviors in the Indonesian Construction Industry. Sustainability. 2023; 15(6): 5261. doi: 10.3390/su15065261
56. Vijaya CJ. A Framework for Implementing Machine Learning algorithms using Data sets. International Journal of Innovative Technology and Exploring Engineering. 2019; 8(11): 155-160. doi: 10.35940/ijitee.k1263.0981119
57. Babi C, Rao MV, Rao VV. Mining Frequent Patterns from Big Data Sets using Genetic Algorithm. International Journal of Engineering Research and Technology. 2018; 11(2): 287-306.
58. Djenour Y, Djenour D, Lin JC, Belhad A. Frequent Itemset Mining in Big Data with Effective Single Scan Algorithms. IEEE Access. 2016; 6: 68013–68026. doi: 10.1109/ACCESS.2018.2880275