Space Constrained Fast Association Rule Mining with Optimal Support **and Confidence Threshold Using** Grammatical Evolution:

An Effective Nudge in Policymaking

Tina Susan Thomas*
Assistant Professor; KCG College of Technology Chennai;
Email: tinaparayil@gmail.com

V. Balaji
Associate Professor; KCG College of Technology Chennai;
Email: balaji.ece@kcgcollege.com

Iranian Journal of Information Processing and Management

Received: 25, Jun. 2021 | **Accepted: 23, Nov. 2021**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Special Issue | Autumn 2022 | pp. 43-64
https://doi.org/10.35050/JIPM010.2022.041

Abstract: In the world of big data and social-media-headed governance and policymaking, data analysis is judged based on the speed and accuracy of execution. This study attempts to modify the existing Association Rule Mining (ARM) techniques by improving the space constraints. Although most of the ARM research is primarily focused on computational efficiency, it has not considered the identification of either the optimal support or the confidence value. Selection of ideal support, as well as confidence value, is vital for the 'ARM's quality. However, with the large dataset availability, the space vector poses the latest challenge in processing. Identification of the optimal parameters adapted to the space model is non-deterministic in nature. This research will focus on a Grammatical Evolution (GE) Association Rule Miner (GE-ARM) to identify the optimal threshold parameters for mining quality rules. Simulations are done using the FoodMart2000 dataset, and then, the proposed method is compared against the Apriori, the Frequent Pattern (FP) growth, and the Genetic Algorithms (GA). Simulation results exhibit substantial enhancements in space and rules generated together with time complexity. Compared to Apriori and FP-tree methods, the proposed GE-ARM achieves lesser runtime by around 20%.

* Corresponding Author

Such an improvisation would categorically change the dynamics of social media analytics by reducing the space constraints and can have more significant ramifications in policymaking. Therefore, such an improvement is undoubtedly an effective nudge in policymaking.

Keywords: Association Rule Mining (ARM), Social Media, Apriori Algorithm, Frequent Pattern (FP) Growth, Genetic Algorithms (GA), and Grammatical Evolution (GE)

Introduction

Advancements in information technology have aided data mining and its techniques in accomplishing the essential objective of extracting business 'individuals' useful information from the dataset in a secure manner (Raj, Vijay & Mahalakshmi, 2016). The basic data mining procedure will involve the revelation of hidden as well as remarkable patterns from the tremendous volumes of data kept in data warehouses or other repositories (Sotudeh et al., 2021). The size of data may be greater than terabytes. Data mining is also inclusive of the integration of techniques from a variety of disciplines like information retrieval, machine learning, database technology, neural networks, statistics (Fesharaki, Shirazi & Bakhshi, 2011). The Knowledge Discovery in Databases (KDD) techniques can extract fascinating patterns within a reasonable time. For pattern extraction to the users, the following steps will constitute the KDD procedure: cleaning, selection, transformation, pre-processing, data mining, and pattern identification (Al-Maolegi & Arkok, 2014).

The key components of a data mining 'system's architecture are a data warehouse, database, a server that will fetch the appropriate data on the basis of the 'user's request, a knowledge base that is utilized as a search standard based on the constraint, and a data mining engine which is the analysis algorithms. In order to find interesting patterns, the pattern assessment unit will interact with the data mining units. Eventually, the graphical user interfaces will enable the user to interact with the data mining unit. Association Rule Mining (ARM) (Boutorh & Guessoum, 2014) can identify correlation relationship associations amongst a huge set of data items. Since tremendous volumes of data are continually gathered as well as stored in the databases, most industries are keen on ARM

for marketing and improving services. This procedure will analyse the customers' buying habits by identifying relations between the different items in the 'customers' shopping baskets. Such 'associations' detection will aid the retailers in developing marketing tactics through insights into the items that are often bought at the same time by their customers.

Most algorithms will operate in two distinct phases in the standard ARM (Kaushik et al., 2020). The first phase will involve the identification of all the frequent itemsets, while the second phase will involve the drawing up of the rules. Apriori and Frequent Pattern (FP)-growth are popular techniques. Normally, these algorithms are treated as the standard ARM algorithms. The standard ARM algorithms will only work with binary data items and lack support for numerical data items. Hence, when the data is in a numeric format (for example, age, height), the data items must be converted from the numerical discretization procedure to the categorical discretization procedure. Numerical Association Rule Mining (NARM) is the term used for defining this procedure of identifying association rules within numerical data.

With the datasets becoming more complex as well as much larger, the standard algorithms such as Apriori will often have to confront problems of high computational complexity during the generation of association rules. For overcoming computational complexity, novel stochastic population-based algorithms are devised for the rule mining procedure as an optimization procedure through the application of search heuristics to the latent optimization problem (Doroudi & Jamshidi, 2021). In general, the heuristics for ARM will involve either single or multiple more interesting measures depending upon whether it will approach as either a single or multi-objective problem. A fitness function will guide the optimization procedure in the search space to improved solutions, where there can be the identification of more association rules having better quality (Ganghishetti & Vadlamani, 2014).

The two distinct groups of stochastic nature-inspired algorithms are Swarm Intelligence (SI) and Evolutionary (EAs) based algorithms. These are population-based approaches that yield new solutions through the application of appropriate variation operators (for example, crossover, mutation, and so on). Some of the various EA-based and SI-based are single and multi-objective methods. This

work has given the proposals for the Genetic algorithm (GA) and the Grammatical Evolution (GE) algorithms for fast ARM algorithm. The rest of this investigation has been arranged into the following sections. Section Two examines the related literary works. Section Three describes the various methods employed in this work. Section Four discusses the experimental results, and Section Five concludes the work.

Literature Review

Thurachon & Kreesuradej (2021) had proposed a novel incremental ARM for frequent itemsets with the utilization of an Incremental Conditional Pattern tree (ICP-tree), Fast Incremental Updating FP Growth (FIUFP-Growth) algorithm, as well as a compact sub-tree feasible. The proposed method recalls the earlier mined frequent itemsets (Sathyanarayanan & Krishnamurthy, 2018) with their support counts. Afterward, this algorithm was used for efficiently mining from the updated database and ICP-tree, thus reducing the original 'database's number of rescans. It was observed from the results that, at a 3% minimum support threshold, the average execution time for pattern growth mining of the algorithm had performed 46% faster in comparison to the FP-Growth, the FUFP-tree, the Pre-FUFP, and the FCFPIM. A new algorithm for multitude-targeted mining, referred to as Guided Frequent Pattern-Growth (GFP-Growth), was recently established (Shabtay et al., 2021). This algorithm was designed for swiftly mining a given set of itemsets with the utilization of a minimal amount of memory. This work had proved that the GFP-Growth algorithm had yielded the exact frequency counts for each itemset of interest.

Moreover, it also showed that the proposed GFP-Growth algorithm had boosted the performance for various problems which required itemset mining. It was also suggested that an Enhanced Apriori Algorithm (EAA) using the context ontology approach for Sequential Minimal Optimization (SMO) could improve the previous algorithms (Sornalakshmi et al., 2020). Ontological knowledge helps to build a hierarchical structure for clustering the items that were composed of ""similar"" concepts, which signified an exact class inside the domain. Every cluster had a defining rule which was on the basis of the relation between the items. The experimental analysis had shown that, compared to semantic ontology, the

proposed method had boosted the accuracy by 2% and minimized the execution time by 25%.

While Kumar & Singh (2019) had chiefly concentrated on the problem of generating association rules for the numeric data. To accomplish this task, the authors took the GA as the basis of this 'problem's solution. The GA was chosen for this task due to its self-improving nature as well as its capability to manage a huge set of solutions. The authors had proposed the GA-based ARM algorithm for the generation of random association rules based on the general property of the dataset. At each iteration, improvement was made on the generated set of rules, and it was further filtered to acquire rules that were better, more interesting, and accurate. The immune mechanism as well as the GA to boost the classical GA was dynamically combined, and also a data ARM method was proposed which was based on the Improved Immune GA (IIGA) for the realization of the Big 'data's effective analysis (Xu et al., 2020).

The simulated outcomes had demonstrated that the proposed IIGA performs better compared to the immune GA and Apriori algorithm with regards to data mining time as well as ARM accuracy and also could be better applied for the analysis of data. The 'research's outcomes had a positive reference significance for the data mining field. When it comes to a large quantum of data, the key to reduce the number of association rules is using Animal Migration Optimization (AMO), known as the ARM–AMO (Chiclana et al., 2018). The proposed algorithm was based on the premise that there would be the deletion of rules that was unnecessary and low support values. At first, there was the application of the Apriori algorithm for yielding frequent itemsets as well as association rules. Later, the AMO was utilized for the reduction of association rules with a novel fitness function incorporating the frequent rules. Experiments proved that, when compared against other relevant techniques, the ARM–AMO had significant reductions in the following: the computational time (for generating the frequent itemset), the memory (for generating the association rule) as well as the number of generated rules.

An evolutionary computing technique for the frequent 'itemsets' retrieval spearheaded by an application of a customized Particle Swarm Optimization (PSO) for mining the FPs through the maximization of the number of frequent

itemsets was identified within the transactional database (Sukanya & Thangaiah, 2020). For management of the frequent 'itemsets' efficient retrieval, the following sequence of techniques was applied: the Standard PSO (SPSO), the Constriction Factor PSO (CFPSO), and the Customized PSO with Local Search (CPSO_LS) method. In the end, an investigation of the proposed techniques was done on the basis of the number of frequent itemsets which were found and the computational time. It was demonstrated from the simulation results that the proposed method had a performance that was superior to that of the PSO 'algorithm's two other variants.

A hybrid algorithm based on the GA and PSO algorithms for the extraction of quantitative association rules would be an improvement (Moslehi, Haeri, & Martínez-Álvarez, 2020). It was possible to acquire accurate as well as interpretable rules by integrating the multi-objective GA algorithm with the multi-objective PSO so as to redress the balance in the tasks of exploitation and exploration. The best rules and the ideal numerical intervals are discovered by the proposed multi-criteria technique where it is not required to find the discretized numerical values or the threshold values. With this proposed method, it was possible to identify the user as well as the most appropriate rules and also the most feasible numerical intervals. The proposed 'method's effectivity was evident from the results, which were acquired from over five real-world datasets.

Ganghishetti & Vadlamani (2014) had formulated the ARM based on the Multi-objective Binary PSO based ARM (MO-BPSO), the hybridized MO-B Firefly Optimization and Threshold Accepting based ARM (MO-BFFOTA), and the hybridized MO-PSOTA based ARM (MO-BPSOTA). While designing, not only metrics like support, confidence, coverage, but also factors like leverage, conviction, interestingness, comprehensibility as well as lift were considered. Afterward, they had utilized them on diverse datasets and also had concluded that the MO-BPSO-TA had the superior performance out of the three.

It is observed from the related works available in the literature that the use of metaheuristic algorithms like GA and PSO significantly improves the ARM compared to Apriori and FP-Growth algorithms. The majority of the works focus on computational efficiency, the space constraint and optimal support, or the confidence value that is threshold values that affect ARM quality are not

addressed. The work attempts at improving the space constraints and optimizing the support and confidence levels. However, with the large dataset availability, the space vector poses the latest challenge in processing. To mitigate this, the binary transformation of data is considered in this work. This research will focus on a Grammatical Evolution (GE) Association Rule Miner (GE-ARM) to identify the optimal threshold parameters for mining quality rules. The use of the genotype-phenotype mapping and Grammars allow the easy integration of domain knowledge; thus, GE is chosen in this work.

Methodology

In the ARM, for a transaction set $D = \{t_1, t_2, ..., t_m\}$, the $I = \{i_1, i_2, ..., i_n\}$ will be a set of all the items. Then X as well as Y will be subsets of items in I, and an association rule [14] will be depicted as follows: X→Y, in which X ⊆ I, Y ⊆ I, X ∩ Y = Φ, X will be the antecedent, Y will be the consequent, and the rule will signify that X implies Y.

Support and Confidence are the two fundamental measures for assessing the association rules. Support refers to the statistical measure depicting the ratio of the records having the 'rule's antecedent as well as consequent. The Confidence gives the percentage of records having the antecedent of a rule which also has that 'rule's consequent in the database. Both these measures can be defined as the below Equations (1 & 2):

(1) $$Support(X \rightarrow Y) = \frac{X \cup Y}{|D|}$$

(2) $$Confidence(X \rightarrow Y) = \frac{Support(X \rightarrow Y)}{Support(X)}$$

The Apriori, FP-Growth and GA-ARM algorithms are used for benchmarking for the proposed GE-ARM.

Apriori Algorithm: a deeper understanding of the quintessential

Apriori algorithm is based on the Boolean association rules of mining frequent itemsets. This 'theory's key principles state that the subsets of frequent itemsets are frequent itemsets, and the supersets of infrequent itemsets are infrequent

itemsets (Shabtay, et al., 2021). The Apriori is employed for finding all the frequent itemsets. In the first iteration, itemset *A* will directly constitute the first candidate itemset C_1. In the k[th] iteration, the candidate's itemset C_k will emerge in accordance with the last 'iteration's frequent itemset L_{k-1}. (Now, the candidate itemset will indicate the potential frequent itemset and also is the *K*-1[th] frequent 'itemset's superset. C_k is made up of *k* frequent itemsets L_k is the expression for an itemset with *k* candidate itemsets.) Afterward, there is the distribution of a counter with a zero initial value to each item set, and also a properly ordered scan of each affair in the database D. When it is ensured that each affair belongs to every itemset, there is an increase in the counter of these itemsets. Upon the scanning of all the affairs, there is the acquisition of the support level in accordance with the actual value of |*D*| as well as the minimum support level of the frequent 'itemset's certain C_k. There is the repetition of the procedure till there are no occurrences of any new item. The connecting step and the pruning step constitute this 'algorithm's two key procedures.

Connecting step: To get L_k, we can connect L_{k-1} with itself and set it to C_k. Assume that the itemsets of L_{k-1} are L_1 and L_2. $L_{i[j]}$ will indicate the j[th] item of *Li*. The assumption is that the affairs, as well as items of the itemset are in order. Conduct the connection $L_{k-1} \lhd\rhd L_{k-1}$, wherein the elements of L_{k-1}, L_1 and L_1, are connectable if they have similar first (*k*-2)[th] items. In other words, the elements of L_{k-1}, L_1 and L_1, are connectable if $(L_1[1] = L_2[1]) \wedge (L_1[2] = L_2[2]) \wedge .....(L_1[k-2] = L_2[k-2]) \wedge (L_1[k-1] = L_2[k-1])$. The requirement of $(L_1[k-1] = L_2[k-1])$ will ensures uniqueness. $L_1[1]L_1[2].....L_1[k-1]L_2[k-1]$ It will be the resultant itemset that connects L_1 and L_2.

Pruning step: C_k will be the superset of L_k. However, all the k frequent itemsets will be contained within C_k. Clear the counters of all the candidate itemsets of C_k in order to ensure L_k and scan the database. Nevertheless, C_k may be huge, and then, there will be a huge number of computations. For decreasing the C_k, the following method will make use of the 'Apriori's traits: any infrequent itemsets having *k*-1 items will not be the subset of frequent itemsets having k items. As a result, if the (*k*-1) subset of a candidate itemset having k items is not in L_{k-1}, the candidate itemset is not frequent and can get removed from the C_k.

Apriori algorithm will employ the bottom-up width search method. When the database of affairs is sparse, this 'database's frequency is often short. Under

this condition, the Apriori algorithm, as well as similar algorithm, are able to get favorable properties. Nevertheless, there will be a sharp drop in the properties when this kind of algorithm meets dense databases (like a population census, telecom, and so on) due to the occurrence of huge amounts of long forms

Frequent Pattern (FP)-Growth Algorithm

Assume that the set of items is $E = \{b_1, b_2, ..., b_n\}$, and the transaction database is *D*. The below equation (3) will provide the expression for the transaction database consisting of multiple numbers of transactions:

(3) $$D = \{R_1, R_2, ..., R_n\}$$

This equation $\{R_1, R_2, ..., R_n\}$ will denote the transactions that constitute the set of items in *E*. Suppose that there is the pattern *P*; it will imply that the pattern '*P*'s set of items, as well as support. The pattern *P* is referred to as an FP it is equal to or more than the predefined threshold *t* of transactions. Definition for the problem of frequent data mining is given as the procedure of finding the whole set of FPs within a database. The FP-Growth algorithm has the two following steps: (1) the FP tree generation and (2) mining the FP with FP-Tree utilization.

FP-Growth (Menaga & Saravanan, 2021) is an ARM algorithm that will identify FPs from the database without any candidate generation, and also it will employ the technique of divide-and-conquer. Initially, it will determine the frequent item list and also will assemble the frequent items in descending order on the basis of their support. Later, it will construct the FP-Tree, and conduct mining on the FP-Tree to generate the sub-database. This procedure is done in an iterative manner. The selection of FP-Growth has benefits like: it will generate the FP-Tree, which has the 'database's complete details despite its smaller size in comparison to the original database, and also it will need less cost for conducting the mining. The pattern growth method employed by this algorithm will minimize the cost required for the 'candidates' generation as well as testing. The technique of divide-and-conquer is employed for minimization of the FP-'Tree's size. This technique will divide the longer FPs into smaller patterns and also will link together the frequent 'items' suffixes. Hence, there is also minimization of the time required to search the FPs.

FP-Tree generation is the FP-Growth 'algorithm's initial step. The FP-Tree must constitute the following traits:

a. The FP-Trees will contain a single root node (null), a number of item prefix subtrees, as well as the frequent item header table.

b. (1) Item-name, (2) count, and (3) node-link will be the three distinct fields of each node in the item prefix subtree. The item-name field will denote which item is represented by the node represents; the count field will denote the total number of transactions which were done to arrive at the node, and the node-link field will get employed for linking one node to the other node within the FP-Tree as well as to take on either the identical name or null during an 'item's absence.

c. Item-name and head of node-link are the two distinct fields of all the frequent item header table entries. The head of node-link will point to the FP-'Tree's initial node and will take on the item name.

The procedure of FP mining is made simpler by the FP-'Tree's traits like the prefix path property and the node-link property.

Genetic Algorithm (GA) Based ARM

The heuristic search algorithm, GA, draws its inspiration from Charles 'Darwin's proposed natural selection procedure. This proposed algorithm is applied for generating feasible solutions and to identify a global optimum within the search area. The solutions need to be encoded at the initialization of the FA. The key operations in GA are (Sarkar, Lohani & Maiti 2017):

(a) Selection - Generation of an *initial population* at the 'procedure's start. The newer 'generations' selection will be carried out based on the individual fitness;

(b) Crossover - will yield the 'population's next generation based on the fitness, which is assessed with the fitness function. The 'solution's quality is measured with this *fitness function*. The average fitness will be greater than the previous generation since the earlier 'generation's best chromosomes have been employed;

(c) Mutation – helps to maintain the genetic diversity amongst diverse generations. In the mutation operation, the solution may experience a complete alteration from the earlier solution. This step is able to mitigate the problem of solutions getting stuck in the local minima at some point.

Upon utilization of the optimization techniques, a general Apriori algorithm will be employed for rule generation. For rule optimization, one of the best techniques is GA. The GA will generate the stronger ruleset so as to optimize the ruleset and also to design a new fitness function that employs supervised learning (Sharma & Tivari, 2012; Haldulakar & Agrawal, 2011).

GA will not directly work on the raw data; there has to be the encoding of the complete data in Binary format (that is, 1 and 0). The below equation (4) will express the 'GA's most critical part, the Fitness 'Function's design:

$$f(x) = \frac{Support(x)}{Min\sup port} \quad (4)$$

Support will denote the Support of New rules generated. Normally, if the q value (Support (x) < min support) will be excluded. It will employ the class learned classifier for predicting which values near to the Maximum value would end up getting rejected.

The q class's value is split into two distinct parts: Q1 and Q2.

q = {Q1, Q2}

Q1 = {Data minsupport < 0.5}

C2 = {Data minSupport > 0.5}

Now, in the equation (5):

$$f(\mathrm{q}) = \frac{Support(\mathrm{C}2)}{Min\sup port} \quad (5)$$

The selection uses the individual fitness, and the concentration pi will indicate the probability of choosing an individual whose fitness value > 1, and f (α) will indicate the value of an individual whose fitness is lower than 1 but near the value of 1.

Now, in the equation (6):

$$pi = \frac{f(xi).e^{-\alpha f(\alpha)}}{\sum f(xj)} \quad (6)$$

Here, α will denote an adjustment factor.

Rules extraction: Generation of the frequent rules will be as per the fitness function as well as the genetic operators. For the eventual mining of strong association rules, there must be a repetition of these 'rules' extraction. The extraction criteria

are as follows: output the rule that fulfills the user-defined minimum confidence; otherwise, abandon the rule.

Proposed Grammatical Evolution (GE) Based ARM (GEARM)

The GE uses linear genome like GA and the mapping from genotype to phenotype uses the rules of grammar in Backus Naur Form (BNF). In GE, the processes takes place at the chromosomal level and not at phenotypic level. The GE-ARM 'algorithm's proposal is to acquire association rules which are independent of any domain or problem. This algorithm will employ GE for defining interpretable individuals. The definition of these individuals is given by using Context-Free Grammar (CFG). There is the utilization of a Backus Naur Form (BNF) grammar to offer technical details for creating the association rules with GE (Boutorh, & Guessoum, 2014).

For the integration of GE with the ARM, it will adapt the GE procedure so as to permit the automatic generation of valid rules. The grammar specifies the antecedents as well as the consequent of the rules in such a way that they are harmonious with the data. Some of the advantages of GE are: it makes use of the genotype-phenotype mapping; Grammars also allow the easy integration of domain knowledge into the search and limit all the possible combinations of components to only those that generate a valid solution.

Grammar

The grammar rules are given by A =⇒ B, where A has only non-terminals and B can be a combination of terminals and/or non-terminals. With the application of the association rules, the non-terminals will ultimately get replaced by the terminals. In more formal terms, the definition of a CFG is given as a quadruple (S, N, T, P), in which S indicates the start symbol, N indicates the non-terminal symbol set, T indicates the terminal symbol set, and P indicates the production ruleset.

The set of variables, as well as their values, will represent the association 'rule's antecedent part. The 'rule's (class variable) consequent can take one of the two values, either positive '"1' (for case) or negative '"0' (for control) states. There is an association of every individual with a case/control. All the elements having a

static form, which indicates that they will not get substituted, will be regarded as terminals (O'Neill & Ryan, 2001).

The GE-ARM Process

The transaction data is converted into binary data to facilitate quick computation of support and confidence.

Below are the GE-ARM 'procedure's various steps:

◇ The GE-ARM 'procedure's training step will commence with the generation of an initial population having N random individuals. In this population, each individual is encoded as an integer-valued vector. The encoding is in the form of string encoding in which each value represents a different item name.

For example, for a transaction of X→Y (item 1, item 3 → item 7, item 8, item 9) is represented as 1 3 0 7 8 9.

◇ The genotype-to-phenotype mapping procedure will make use of the above grammar and always will commence with the Start symbol. If the 'genome's end has been reached and the mapping procedure continues to be unfinished, then the genome will get wrapped over, and the integers will get read again from the initial vector. This wrapping procedure will continue T times (predefined value) or if all the non-terminals have been substituted, then there is the mapping procedure is terminated.

◇ The resultant output string will then determine the set of N association rules in which there is mapping of every individual in the initial population to an association rule. There is an assessment of every association R on the training set, and its fitness will get recorded. The objective of the fitness function is to maximize the support and confidence, thus obtaining associations of greater strength.

◇ There is a selection of the best N-rule solutions for crossover as well as reproduction. The operations of crossover and mutation are done at the chromosomal level (the integer-valued vector) and not at the level of the association rules. The resultant new generation, which has the best rules and has the same size as the original population, will get iteratively utilized in the cycle till the fulfilment of certain criteria, after which the GE-ARM procedure will come to a halt. This criterion of termination can be either a limit on the

number of generations or a zero classification error.

◇ After every generation, there is the identification of the best solution. At the GE-ARM 'evolution's end, the overall best solution will get picked as the optimal AR set.

The flowchart of the proposed GE-ARM is shown in Figure 1.

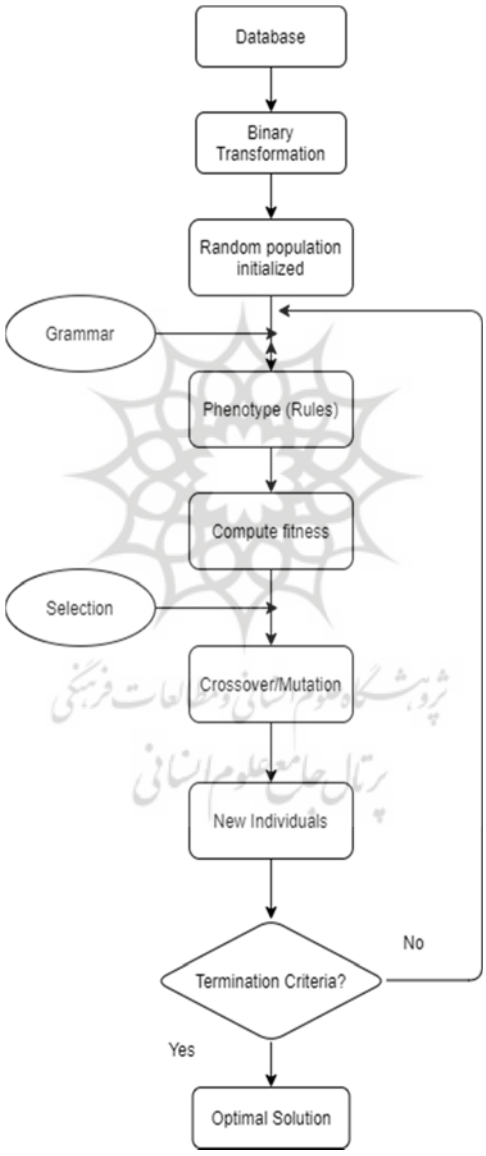

Figure 1. Flowchart of GE-ARM

Every generated rule will indicate a potential transaction amongst the instances, and the final output will indicate a list of transactions. GE-ARM has a parameter set that must be initialized. Upon initialization, the data is split into 10 equivalent parts for 10-fold cross-validation. While 9/10 of the data will get employed for training, the remainder 1/10 of the data will later get employed for assessment of the 'model's predictive capability.

Dataset

For the purpose of evaluation, there is the utilization of Foodmart, a retail 'store's real-life sparse dataset, with real utility values that are acquired from the Microsoft Food-mart 2000 database. In this data table, the number of product items is 1,560. For effective mining of meaningful association rules, this experiment will differentiate the products as per the product category given by the data table. Hence, there is a classification of the products into 34 categories, wherein each category has a corresponding product category id. With regard to the data selection, there is the random selection of 6,000 customers together with their corresponding transaction data at distinct times. After the arrangement, these 6,000 customers will have a total of 12,100 transaction records.

Results and Discussion

The experiments were conducted on the Matlab platform. The Weka tools was used to code the optimization techniques. The parameters of the algorithm were set as follows: for population size = 50 individuals; chromosome length ≥ 4, generation size = 250; crossover rate = 0.9; mutation rate = 0.1. The Apriori, FP-tree, GA and proposed GE methods are evaluated using Foodmart2000 for various support thresholds. In this section, the runtime, memory utilization and percentage increase as shown in Figures 2 to 4.

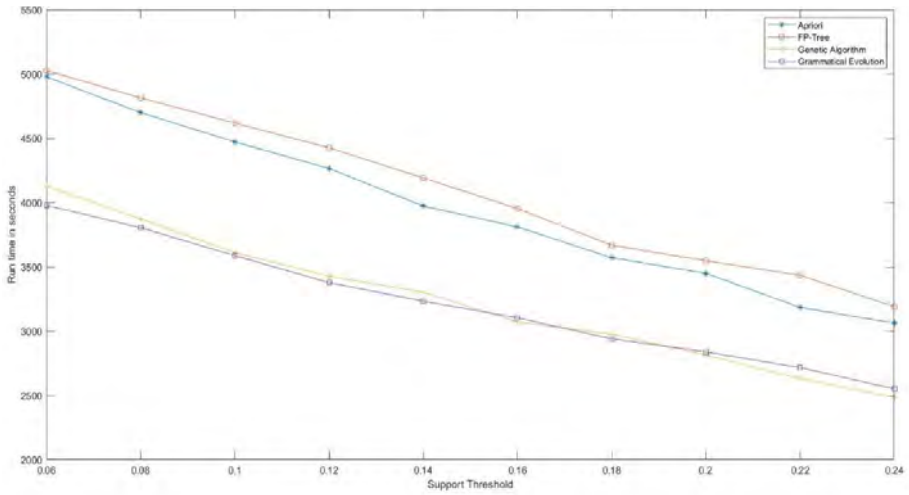

Figure 2. Runtime

From Figure 2, it can be observed that the proposed GE achieves lesser runtime by around 20% compared to Apriori, FP-tree at various support thresholds, respectively. It is observed that both GA and GE methods have similar runtime, as it is not required to subjectively set up the threshold values for minimal support and confidence, computation time is saved.

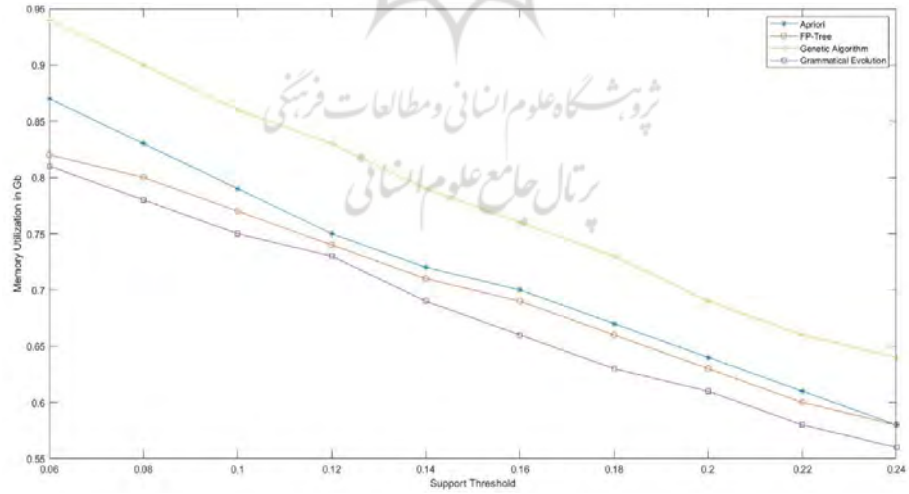

Figure 3. Memory Utilization

From Figure 3, it can be observed that the proposed GE has lesser memory utilization than Apriori, FP-tree and GA when compared with various support thresholds, respectively. Though GA has similar runtime to the proposed GE, the memory required is significantly higher for GA. The use genotype and phenotype in GE helps in forming better solutions.

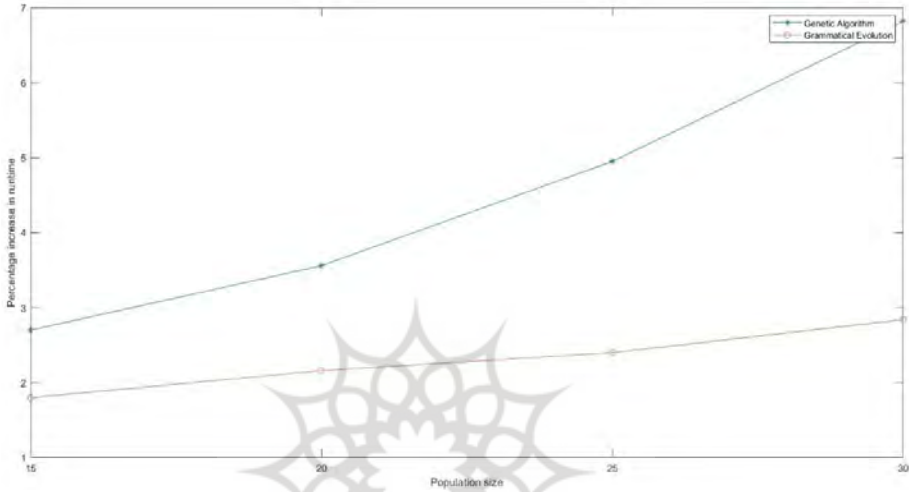

Figure 4. Percentage Increase in runtime

From Figure 4, it can be observed that the proposed GE requires lesser time for execution than GA as the population size increases. The proposed GE achieves fast association rule mining compared to existing algorithms.

Implications to social media and policy making

Data management has been the primary concern for any entity. This conundrum is more pronounced in circumstances where there is uncontrolled flow of data and information through the different channels of social media (Tanantong, & Ramjan 2021). In such situations, brevity and speed act as the determining factors. Many a time the accountability of social media is discredited due to the systemic flaws of data analysis and management which mainly corresponds to the time and space constraints. This study categorically underscores a mining methodology which can have significant effect on the critical understanding of the transactions in the social media. This can effectively facilitate the targeted policymaking decisions mainly

by assimilating and deciphering individual to individual interaction. Therefore, this study comes up as an effort to change the decision pattern analysis for policy making and good governance. It can also be instrumental in embellishing national security by effective target-oriented policy executions as a result of fast and efficient transaction analysis.

Conclusion

The popular ARM is a well-researched method for the detection of fascinating relationships between variables in data 'mining's large databases. In general, the association rules are essential for the parallel satisfaction of user-specified minimum support as well as user-specified minimum Confidence. The Apriori algorithm will continuously generate candidate itemsets and will employ minimal support and minimal confidence for filtering these candidate itemsets to identify the high-frequency itemsets. Initially, the FP-Growth algorithm will extract the FPs from the database via the support 'threshold's application and also will generate the association rules from the FPs via the confidence 'threshold's application. The GA is a search heuristic is effectively used for ARM.

These algorithms are vital for the discovery of association rules since they work with the global search for the detection of a set of item frequency, and also, they are much simpler than the other commonly-used data mining algorithms. This work has given the proposal for a new approach, called the GE-ARM, which will utilize the GE to discover an association rule set. It is evident from the simulation results that, when compared with diverse support thresholds, the proposed GE has better performance (runtime as well as memory utilization) than the Apriori, the FP-tree, and the GA.

References

Al-Maolegi, M., & Arkok, B. (2014). An improved Apriori algorithm for association rules. arXiv preprint arXiv:1403.3948.

Boutorh, A., & Guessoum, A. (2014). Grammatical Evolution Association Rule Mining to Detect Gene-Gene Interaction. In *BIOINFORMATICS* (pp. 253-258). In Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS-2014), pages 253-258 ISBN: 978-989-758-012-3

Chiclana, F., Kumar, R., Mittal, M., Khari, M., Chatterjee, J. M., & Baik, S. W. (2018). ARM–AMO: an efficient association rule mining algorithm based on animal migration optimization. *Knowledge-Based Systems, 154*, 68-80.

Doroudi, F., & Jamshidi, Z. (2021). Assessing the Components of Information Security in Accessing & Use of Digital Libraries. *Iranian Journal of Information processing and Management, 37*(1), 117-134.

Fesharaki, M., Shirazi, H., & Bakhshi, A. (2011). Knowledge Acquisition from Database of Information Management and Documentation Softwares by DataMining Techniques. *Iranian Journal of Information processing and Management, 26*(2), 260-283.

Ganghishetti, P., & Vadlamani, R. (2014). Association rule mining via evolutionary multi-objective optimization. In International Workshop on Multi-disciplinary Trends in Artificial Intelligence (pp. 35-46). Cham: Springer.

Haldulakar, R., & Agrawal, J. (2011). Optimization of association rule mining through genetic algorithm. *International Journal on Computer Science and Engineering (IJCSE), 3*(3), 1252-1259.

Issac, A. C., & Baral, R. (2020). A trustworthy network or a technologically disguised scam: A biblio-morphological analysis of bitcoin and blockchain literature. Global Knowledge, Memory and Communication.

Kaushik, M., Sharma, R., Peious, S. A., Shahin, M., Yahia, S. B., & Draheim, D. (2020, November). On the Potential of Numerical Association Rule Mining. In International Conference on Future Data and Security Engineering (pp. 3-20). Springer, Singapore.

Kumar, P., & Singh, A. K. (2019). Efficient generation of association rules from numeric data using genetic algorithm for smart cities. In Security in Smart Cities: Models, Applications, and Challenges (pp. 323-343). Cham: Springer.

Menaga, D., & Saravanan, S. (2021). GA-PPARM: constraint-based objective function and genetic algorithm for privacy preserved association rule mining. *Evolutionary Intelligence*, 8(2) 1-12.

Moslehi, F., Haeri, A., & Martínez-Álvarez, F. (2020). A novel hybrid GA–PSO framework for mining quantitative association rules. *Soft Computing, 24*(6), 4645-4666.

O'Neill, M., & Ryan, C. (2001). Grammatical evolution. *IEEE Transactions on Evolutionary Computation, 5*(4), 349-358.

Raj, B. B., Vijay, J. F., & Mahalakshmi, T. (2016). Secure data transfer through DNA cryptography using symmetric algorithm. *International Journal of Computer Applications*, *133*(2), 19-23.

Sarkar, S., Lohani, A., & Maiti, J. (2017, March). Genetic algorithm-based association rule mining approach towards rule generation of occupational accidents. In International Conference on Computational Intelligence, Communications, and Business Analytics (pp. 517-530). Singapore: Springer.

Sathyanarayanan, D., & Krishnamurthy, M. (2018). Association Rule Mining Using Frequent Itemsets Generation by Anti-Mirroring Of Bit Vectors. *International Journal of Pure and Applied Mathematics*, *118*(20), 5033-5044.

Shabtay, L., Fournier-Viger, P., Yaari, R., & Dattner, I. (2021). A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data. *Information Sciences, 553*, 353-375.

Sharma, A., & Tivari, N. (2012). A survey of association rule mining using genetic algorithm. *International Journal of Computer Applications & Information Technology, 1*(2), 5-11.

Sornalakshmi, M., Balamurali, S., Venkatesulu, M., Krishnan, M. N., Ramasamy, L. K., Kadry, S., ... & Muthu, B. A. (2020). Hybrid method for mining rules based on enhanced Apriori algorithm with sequential minimal optimization in healthcare industry. *Neural Computing and Applications*, 13(3) 1-14.

Sotudeh, H., Yousefi, Z., Khunjush, F., & Ghanbari Aloni, F. (2021). Content analysis and Opinion mining of Tweets about Open Access and its Main Features. *Iranian Journal of Information processing and Management, 37*(1), 305-329.

Sukanya, N. S., & Thangaiah, P. R. J. (2020). Customized Particle Swarm Optimization Algorithm for Frequent Itemset Mining. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-4). IEEE. Sweden.

Tanantong, T., & Ramjan, S. (2021). An Association Rule Mining Approach to Discover Demand and Supply Patterns Based on Thai Social Media Data. *International Journal of Knowledge and Systems Science (IJKSS)*, *12*(2), 1-16.

Thurachon, W., & Kreesuradej, W. (2021). Incremental Association Rule Mining with a Fast Incremental Updating Frequent Pattern Growth Algorithm. *IEEE Access, 9*, 55726-55741.

Thomas, T. S., & Issac, A. C. (2018). Real time monitoring of the health of infants. https://aisel.aisnet.org/amcis2018/TREOsPDS/Presentations/98/

Xu, B., Ding, S., & Li, Y. (2020). Data association rules mining method based on genetic optimization algorithm. *In Journal of Physics: Conference Series 1570*(1), 012006. IOP Publishing.

Tina Susan Thomas

Tina Susan Thomas is an Assistant Professor in Department of Information Technology, at KCG College of Technology, Chennai. She has keen interests in research and her papers were published in Association of Information System (AIS) library and other reputed international journals. She is an active member of Computer Society of India (CSI) and Indian Society for technical Education (ISTE). Her academic and research interests include Knowledge Management, Information system, Datamining and Cyber security.

V. Balaji

Dr. V. Balaji is working as an Associate Professor in the Department of Electronics and Communication Engineering at KCG College of Technology, Chennai. He is having 16 Years of Experience in the teaching filed. He has published more than 50 Papers in the SCI/ WOS/ Scopus and Google Indexed Journals. He has got 4 Patent to his name. His research areas include Wireless Networks, Networks, Image Processing, Data Mining.