International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

An Optimization based Algorithm for Association Rule Mining in Large Databases

K. Kala

Abstract: ***In recent trends, the information and science technology has been developed which makes most of the research in Association Rule Mining (ARM) to be focused on enhancing the computational efficiency.But the searching for most appropriate data in large databases become more difficult.The major difficulties in this is determining the threshold values for confidence and support which causes effect on the quality of the association rule mining. Thus a novel optimization based algorithm for association rule mining is proposed in this work for the purpose of enhancing the computational efficiency and also to determine the appropriate threshold values automatically. Initially the input raw data is taken from the large transactional database.Then this input raw data is preprocessed and converted into binary type data.After this, the Modified Binary Particle Swarm Optimization (MBPSO) algorithm is used which involves five phases namely 1) Encoding – where a string based encoding is followed that encodes every single item into respective string chromosome, 2) Fitness Evaluation – where the 'maximization' has been determined using association rule type parameters, 3) Population generation – where an initial population has been generated using particles with best fitness values, 4) Searching – where best particles are searched in order to prevent it from falling apart from search space whenever the position is updated, 5) Stopping circumstance – as it is essential to complete the particle evaluation, by finding respective confidence and support as the minimal threshold values. This would be helpful in deriving valuable information in ARM. The performance of this proposed methodology is validated and compared with the tradition association rule mining algorithms. This proposed methodology offers better results with increased efficiency and proves its superiority.***

Key words: Data Mining, Association Rule Mining, Modified Binary Particle Swarm Optimization, Confidence and Support Value, Fitness Function.

I. INTRODUCTION

Due to the advancements in the information and computer science technology, enormous amounts of data were gathered and recorded in the system memory. It is possible to understand the behaviors, facts and natural phenomena of the gathered data by analyzing it. This increases the concept of data mining as a significant field[1]. This data mining helps to extract the most relevant as well as useful information from the data which is gathered and convert them into comprehensible form of knowledge for the purpose of further utility. Data mining is comprised of several methods from database systems, artificial intelligence, mathematics, statistics and machine learning. Several data mining techniques are formulated like sequence pattern discovery, association rule mining, etc.The most difficult issue in data mining is to discover the association rules[2]. It is a firm approach for extracting the information from the structured databases, also they offer support for the detection of new, useful and comprehensive knowledge. Normally the data which is used for data mining are represented as primary data are cleaned and preprocessed based on the requirements of the mining algorithms. But this primary data is available only for the short time which are not recorded. They can be processed in real time applications and deleted after obtaining the results. For the large database, this association rule mining and the frequent item set mining becomes the most widely used techniques which are utilized for identifying the data items that are frequently occurred and interesting association relationships amongst them[3]. These techniques are employed in various applications such as web usage mining, market basket analysis, prediction, bioinformatics and health care. It is very easy to extract the association rules after mining the frequent item sets and determining their supports. Depending up on the frequent item set mining approaches, the association rule mining algorithms are developed. There are several traditional association rule mining and frequent item set mining are designed such as Apriori, FP-Growth, Éclat, etc. for the database setting in which the raw data is stored for mining processes. The rules which are used for generating all the possible item sets need to create association rules for frequent item sets. The association rule mining is a rule based learning approach which helps to determine the important relationships amongst the data items in the data set. Among several association rule mining, the most familiar approach is the Apriori algorithm which is the fundamental algorithm of Boolean association rule mining. Moreover an enhanced algorithm which is referred to as FP-Growth algorithm is introduced. Inorder to meet the necessities in the data mining processes, the serial algorithm becomes more difficult due to the data accumulation in a continuous manner. Thus parallel mining algorithm is developed. Depending up on the various situations, the association rules are categorized as follows.

Manuscript published on 30 September 2019
* Correspondence Author
K.Kala Head & Associate Professor, Department of Computer Science,
Nachiappa Swamigal arts and Science College, koviloor, India.
Corresponding author mail id: kalaphdk12@outlook.com

- The association rules can be partitioned ad numeric and Boolean, depending up on the classification of variable that are handled in the rules.
- The association rules are divided into two levels such as single level association rules and multi-level association rules, based on the abstraction levels of data in the rules.
- The association rules are subcategorized as single dimensional as well as multi-dimensional, depending up on the data which is involved in the rules.

Various traditional algorithms are used for association rule mining, still it lacks in some drawbacks such as increased time consumption with computational complexities. Hence a novel algorithm is proposed based on the optimization algorithm for obtaining best association rules in data mining.

The main objectives of this work are as follows:

- To search the best and appropriate association rules from large transactional databases with increased efficiency
- To determine minimal threshold values for support and confidence of the association rule mining
- To determine best association rule using Modified Binary Particle Swarm Optimization.

The remaining sections of this paper are systemized as follows: section II investigates the various traditional association rule mining approaches in data mining. Section III describes the working procedure of the proposed methodology in detailed manner. Section IV demonstrates the performance of the proposed methodology. Finally the conclusion of this propos4ed work is mentioned in Section V.

II. RELATED WORKS

In this section, various tradition approaches that are used for association rule mining in data mining is discussed. Also their working procedure along with their merits and demerits are illustrated as follows. A research had been introduced the most popular algorithms which was referred to as Apriori algorithm. It helped to extract the frequent item sets from large datasets. The knowledge from those large datasets could be discovered by obtaining the association rule. The major drawback of acquiring more time for scanning the entire database searching on the frequent item sets depending up on this algorithm was indicated in this suggested work. An enhancement on Apriori algorithm was presented which reduced the wasted time by scanning some item sets only. The results showed that this enhance Apriori algorithm acquired less time for scanning the large dataset compared to the original Apriori[4]. A research proposed a protocol for the purpose of securely mining the association rules in horizontally distributed datasets. This protocol was used depending up on the fast distributed mining algorithm which was an unsecured distributed version of Apriori algorithm. There were two major ingredients of novel secure multi-party algorithms in this protocol in which the one assessed the union of private subsets and the other tested the addition of a component which was detained by a player in a subset detained by another. The main intention of this research was to devise an effective protocol for set addition verification which helped in the presence of semi honest third party. But the major issue of this protocol was its computational as well as communication costs[5].

A research recommended a new neutrosophicassociation rule algorithm which helped to overcome the problems in traditional association rule mining algorithms such as ineffectiveness, reduced quality of generated association rules because of the sharp boundary issues. This recommended algorithm utilized a new approach for the generation of association rules with the deal of non-membership, membership and indeterminacy function items. This was conducted to an effective decision making system by taking into account of all the imprecise association rules. The experimental analysis was conducted and compared with the fuzzy as well as neutrosophic mining to prove the validity of the recommended system. This approach offered better results with increased number of generated association rules[6]. A new multiobjective evolutionary algorithm named MOPNAR for the purpose of mining a reduced set of positive as well as negative quantitative association rules with minimal computational cost had been designed. The plan of this work was extended a current multiobjective evolutionary algorithm depending up on the decomposition which helped to accomplish the evolutionary learning of the attributes' intervals and the selection of condition for every rule during the introduction of an external population and the process of restart for storing all the non-dominated rules that were found. Also this devised method helped to enhance the diversity of the determined rule set. This method had three major goals such as performance, interestingness and comprehensibility for obtaining the rules that were easy to understand, offered better coverage and interesting of the dataset. The performance of the devised method was evaluated and compared with various real world applications. The results revealed that the devised method offered better scalability with less computational cost even when the size of the problem got increased[7]. A study had been made about the issue of mining fuzzy association rules from the ubiquitous data streams. Here a novel Fuzzy Frequent Pattern Ubiquitous Streams (FFP_USTREAM) technique was developed which combined the concepts of fuzzy with the ubiquitous data streams and employed the sliding window approach for mining the fuzzy association rules. Moreover the effectiveness as well as the complexity of this technique was discussed. The results were more helpful for several practical applications which offered more flexible as well as significant decisions like calculating the required stocks in the applications of retail, treatment methods in medical and precaution methods in road safety fields. But they had some issues due to the window techniques and also incompatible[8]. An effective approach had been developed for the purpose of mining lass association rules (CARs) with the item set constraint depending up on the stricture of the lattice as well as the difference amongst two sets of object identifiers (diffset). Initially all the frequent item sets in the dataset was stored by building a lattice structure.

Then the diffset was utilized rather that the whole set of object identifiers for minimizing the memory usage. The rules were generated which satisfied the item set constraint by traversing the lattice.
The effectiveness of the developed approach was validated and the results demonstrated that this approach outperformed compared to the existing approaches[9].

An outlook of characterizing the minimum supports of elementsets while the elements had minimal supports had been offered. In this work, the algorithm was utilized for collecting the sample association rules which were taken from basic Apriori algorithm with the multiple minimum support using maximum constraints algorithm. This algorithm was implemented and compared with the other algorithms by using new comparison algorithm. This was done for proving the superiority of the proposed algorithm. Moreover the comparisons were carried out on different data groups. From the comparative analysis the results showed that the proposed algorithm offered fast implementation than the other algorithms. Also while implementing this algorithm on dataset provided better accuracy results without affecting the number of rules that were generated[10].

A method by utilizing the averaging techniques in which the Apriori algorithm specified about the minimum support in an automated manner had been developed. The main objective of this method was to accomplish an improvement in Apriori algorithm by trying fuzzy logic for distributing the data in various clusters initially. Later the most suitable threshold was tried to introduce in an automatic manner. The effectiveness of this method was analyzed by using similar database of the comparison approach and the other three algorithms such as Éclat, Apriori and FP-Growth. The results revealed that this method offered better performances compared to the other algorithms[11].

A research dealt with the issues of numerical Association Rule Mining (ARM) by utilizing a multi-objective viewpoint. This could be done by introducing a multi-objective particle swarm optimization algorithm which helped to identify the numerical association rules in a single step only. In this multi-objective optimization approach, several objectives could be identified such as interestingness, comprehensibility and confidence. At last the Pareto optimality was utilized for extracting the best association rules. Also the rough values which was comprised of upper as well as lower bounds were utilized for showing the intervals of attributes. The effect of operators that were used in this approach were analyzed in the experimental analysis and compared with the most familiar evolutionary based approaches for ARM. The results demonstrated that the suggested approach extracted the most reliable, interesting and comprehensive association rules. But this approach was required to get higher support values for association rules[12].

Gibbs sampling induced stochastic search model for randomly sample association riles from the item sets had been developed. The rule mining was performed from the reduced transaction dataset which was generated by the sample. Moreover a general rule significance measure was introduced for the purpose of directing the stochastic search. This helped the randomly generated association rules that were comprised of ergodic Markov chain, the entire most significant rules in the item set could be uncovered from the reduced dataset. The main intention of utilizing the Gibbs sampling based approach was to reduce the item sets for data mining in which the Apriori algorithm was implemented for determining the most significant association rules which were subjected to negligible data loss. The simulation results of the developed model offered better performance[13].

A research had been made where it described about the bacterial colony networks and their skills for exploring the resources as a tool in order to mine the association rules in databases. This algorithm was designated for maintaining the different solution for the issues. This algorithm was inspired by the exploratory nature of the surrounding resources by a colony of bacteria for extracting the association rules of items in the transactional databases. This algorithm had the capability to eliminate the genic conversion issues. The experimental analysis was carried out and compared with two bio-inspired heuristics. Also the most familiar Apriori algorithm was compared. The bacterial colony algorithm offered better results in data mining. But it required some other bio-inspired algorithms for solving the issues in real world data mining and optimization[14].

A trendy solution has been proposed in a research where the issues such as requirement of high storage space for the purpose of saving enormous data which resulted in the generation of frequent item set and the requirement of encoding methods in which individual symbols were utilized for every possible value of an attribute of the item set. Initially the genetic algorithm was utilized for defining the maximal frequent item set which did not required enormous storage. Moreover the genetic algorithm was forced to work directly on the database which avoided the requirement of encoding method. The assessment of the proposed method was done depending up on the recommendation to utilize the variable length individual in the population. This modified algorithm was implemented in real database and the results proved its superiority[15].

A Weighted Fuzzy Privacy Preserving Mining (WFPPM) for extracting the sensitive association rules had been utilized. The sensitive rules could be identified completely by using this approach with the calculation of the weights of rules. Initially the FP-Growth was implemented for mining the association rules from the databases. After this, the fuzzy was executed for determining the sensitive rules amongst the mined rules. The experimental results revealed that this approach extracted sensitive rules by taking into account of the weight of every single parameter instead of depending up on the minimal threshold value of the confidence as well as support. Also this approach had the capability to implement in various applications such as bank and other financial sectors[16].

A data mining algorithm in order to derive the fuzzy temporal association rules had been constructed. Initially every quantitative value was transformed into a fuzzy set by utilizing the given membership functions.

Meanwhile, the lifespans of the item were gathered and recorded in a temporal data table using the process of transformation. After that the algorithm assessed the scalar cardinality of every linguistic term of individual item. Depending up on the counts of the fuzzy as well as the lifespan of the item, the process of mining was performed for determining the fuzzy temporal association rules[17]. The main objective of this work was taking the lifespan of every item into account. This helped the approach to determine more data in a given transactional database compared to the other approaches. But they had some drawbacks in determining the fuzzy temporal mining for items with multiple minimal supports and selection of suitable membership functions for mining.

A novel method of super edge definition for improving the capability of dealing with the issues had been proposed in a research. Also an association rule redundancy processing algorithm depending up on the hyper graph was studied. Here the idea of hyper graph and the system were introduced for exploring the hyper graphconstruction on 3D matrix model. Moreover the association rules were transformed into directed hyper graph by utilizing the association rules redundancy as well as the loop detection depending up on the directed hyper graph and redefined the adjacency matrix. The redundancy detection as well as the loop were converted into the connected blocks and circles processing in hyper graph. It offered a novel concept and procedure for redundant processing of association rules. For the purpose of practical projects, the novel method was implemented and the results offered better results which helped to determine the knowledge with high quality from the high dimensional data[18].

III. PROPOSED WORK

This section demonstrates the working procedure of the proposed methodology in a detailed manner. The flow of the proposed association rule mining approach is depicted in Fig.1. This approach is comprised of two sections such as preprocessing and mining. The first section offers the processes of fitness value assessment of the particle swarm. Hence the data can be converted and recorded in the binaryformat. After this the IR value is calculated for setting the searching range of the particle swarm. The major contribution of this proposed methodology is in the second section which employed the Modified Binary Particle Swarm Optimization algorithm for the purpose of mining the association rules. Initially the encoding process is carried out in particle swarm which is similar to the encoding process of chromosome in the genetic algorithm. Then the population of particle swarm isgenerated depending up on the assessed fitness value. Finally the MBPSO searching process is done till the best particle is obtained. The calculation of support as well as confidence of the best particle denotes the minimal support as well as minimal confidence. This further helps in association rule mining.

A. *Preprocessing*

Initially the input transactional data from the large database is taken and preprocessing is carried out. In this the input raw data is converted into binary type data and every data is recorded as 0s or 1s. Here the database scanning operation is done in this approach and measures the support and confidence value in an effective as well as easiest manner. Also the process of noise removal, removing the white spaces, special characters, symbols and other irrelevant items from the input data is done in this preprocessing step. Algorithm for preprocessing is described as follows.

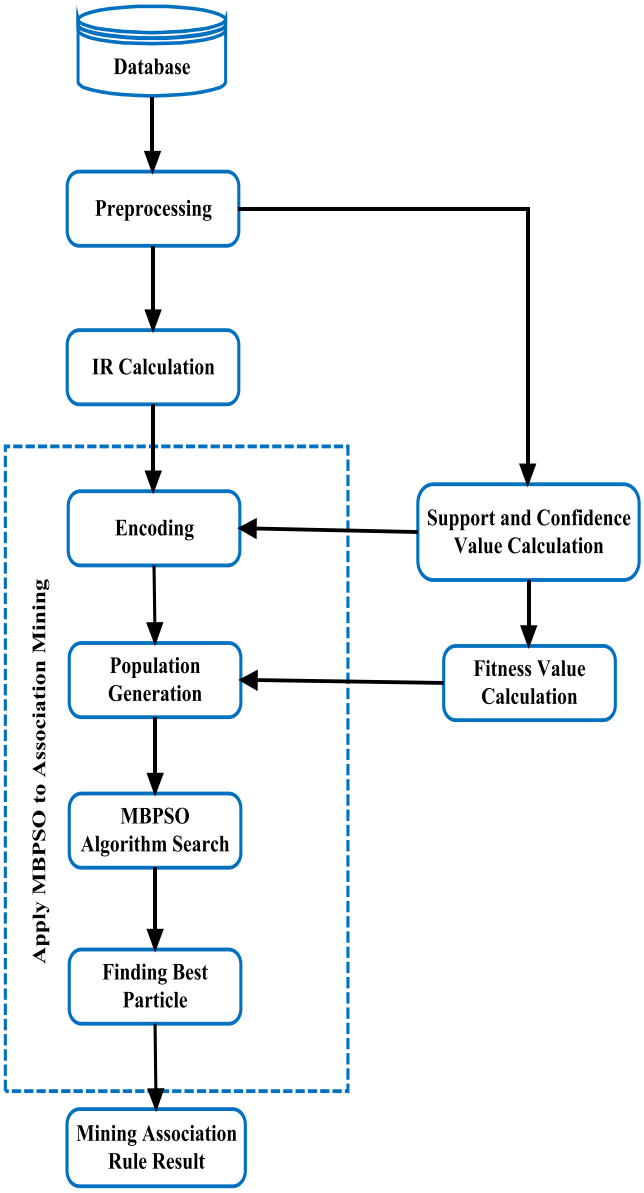

Fig. 1.Flow of the proposed Methodology

Algorithm I: Preprocessing

Input:Input Data Set
Output:Pre-processed Data
Procedure:
#Pre-Processing the data.
Step 1: Noise Removal and Removing special characters
Step 2: while line! = null
Step 3: alphabets = line split by ","// Split by commas
Step 4: for alphabet : alphabets //for each alphabets is converted into lower case
Step 5: for i=0 to length (alphabet)
Step 6: for i=0 to length (String)
Step 7: if 48<String<58
Step 8: sum=sum+str.character-48
Step 9: End if
Step 10: End for
Step 11: while (sum>0) //Iterating sum values to sum up each position
Step 12: temp = sum%10
Step 13: sum1=sum1+temp
Step 14: sum = sum/10
Step 15: End while
Step 16: End if
Step 17: End while.

After this the calculation of IR values is carried out. The main intention of this calculation to generate more significant association rules. Also this IR calculation helps to improve the efficiency of the search for deciding the rule length that is generated by the chromosomes in the particle swarm. The search for irrelevant item sets are avoided by this IR calculation and it is given by,

$$IR = \left[log\left(aTN(a)\right) + log\left(bTN(b)\right)\right]\frac{T(a,b)}{Total\ T} \quad (1)$$

Where, a and b denotes the length of the item set,
$TN(a)$represents the number of transactions records that contains a and b products,
$T(a,b)$denotes the number of transaction records that are purchased a to b products.
$Total\ T$represents the total number of transactions.

B. MBPSO algorithm

After preprocessing step, the implementation of MBPSO algorithm to the association algorithm takes place. It is the major part of this proposed methodology which uses MBPSO for extracting the best fitness value. This algorithm involves encoding, fitness values evaluation, population generation, searching best particle and stopping condition. The steps that are involved in MBPSO algorithm and the generation of association rules are discussed as follows:
In this proposed methodology, there are five steps involved which are listed as follows:

- Encoding
- Fitness value evaluation
- Population generation
- Searching of best particle
- Stopping circumstance.

1. Encoding

Depending up on the association rule mining definition, the intersection of the association rule of A to B item set$A \rightarrow B$ should be empty. This can be defined as the items which appear in item set A should not appear in item set A and vice versa. This both the front as well as back propagation is used for chromosome encoding. The item set which is before the front splitting point is referred to as "item set A", whereas the item set amongst the front splitting as well as back splitting is referred to as "item set B". Here the approach used for chromosome encoding is the "string encoding" in which the individual value denotes a dissimilar item name. The representative value of individual value of every single item can be encoded into string type chromosome by the respective order. Then the IR value which is assessed in the previous step can be utilized for selecting the front as well as back splitting points of the chromosomes.

2. **Fitness value evaluation**

The significance of an individual particle is evaluated by using the fitness value which is comes from the fitness function. Here the target function is employed for determining the fitness function value and this can be expressed as

$$Fitness\ (m) = confidence\ (m) X \log\left(support(m) X\ length(m) + 1\right) \quad (2)$$

Where, $Fitness\ (m)$ represents the fitness value of association rule type m,
$confidence\ (m)$denotes the confidence of the association rule type m,
$(support(m)$symbolizes the actual support of the association type m,
$length(m)$denotes the length of the association rule type m.
The main motive of this fitness function is maximization. The strength of the association becomes greater when the confidence as well as the particle support is larger which stated that the association is a significant rule. Before evaluating the fitness value, the confidence, support and length should be calculated. This proposed methodology utilizes the binary type data search method in which the original data is first arranges into two dimensional matrix where each row denotes the data records and each column denotes the product items. Here the searching process is carried out in column by column which reduces the search time and increases the efficiency of the evaluation in a great manner.

3. Population generation

It is essential to generate the initial population for applying the evolution process of MBPSO algorithm. Here the particles which are having larger fitness values are selected as population. In this population, the particles are represented as initial particles.

4. Searching best particles

The particle which has the maximum fitness value in the population is initially selected as the "gbest" and the initial velocity of the particle can be set as $\vartheta^0 = 0$. The initial position of the particle is denoted as "pbest" and this can be updated.

The calculated values are integer, thus a method for constraining the search is designed. This constrained method is used for assessing the distance amongst the new position of the particle and all the particles in the range before the position of the particle is updated. The particles which have the smallest distance can be selected and noted as the new position of the particle.

The distance can be assessed by using Euclidean distance method and it is given by,

$$dist(a^x, b^y) = \sqrt{\sum_1^d \left(a_i^x - b_i^y\right)^2} \quad (3)$$

Where, a^x denotes the position of the particle at xth update, b^yrepresents the particle number y in the constrained range, d symbolizes the dimension of the search space

The closest particle can be selected as the new position of the target particle. This helps to prevent a particle from falling apart from the search space when the position is updated.

5. Stopping circumstance

The design of stopping circumstance is essential for completing the particle evolution. Here the process of evolution is stopped when the fitness values of all the particles are fixed. After every 100 iterations, the stopping circumstance is occurred and the evolution process of particle swarm is accomplished. After finding the best particle, the corresponding confidence as well the support are suggested as the value of minimal confidence as well as the minimal support. This helps to extract the valuable information in the association rule mining. The algorithm for MBPSO is explained as follows.

Algorithm II: MBPSO Algorithm

Input:Featured Attributes
Output:Selected Attributes.
Procedure:
Step 1: //Initializing random particles
Initialize some random values to the variables
double w = 0.9, c1 = 2.05, c2 = 2.05, r1 = 0.0, r2 = 0.0, xMin = -5.12, xMax = 5.12, vMin = 0, vMax = 1, wMin = 0.4, wMax = 0.9, phi = c1+c2, chi = 2.0/Math.abs(2.0-phi-Math.sqrt (Math. pow (phi, 2)-4*phi)), nInfinite = Double.NEGATIVE_INFINITY, gBestValue = nInfinite

intNp = 569,Nd = 10, Nt = 1

double[] pBestValue = new double[Np], gBestPosition = new double[Nd],tFitnessHistory = new double[Nt], M = new double[Np]; double[][] pBestPosition = new double[Np][Nd], R = new double[Np][Nd], V= new double[Np][Nd]

Step 2: While line1 = null // loop until null
Step 3: double rand=0
Step 4: str[][] = line.split ",”//split by commas
Step 5: For → I to str.length // Iterating each alphabets
Step 6: String str=str[]
Step 7: rand = double.parseDouble(str) //Converting into double
Step 8: For → p to Np //for each best position
Step 9: For → I to Nd // for each best value
Step 10: R[p][i] = xMin + (xMax-xMin)*rand //calculating R
Step 11: V[p][i] = vMin + (vMax-vMin)*rand //calculating V
Step 12: If rand < 0.5
Step 13: V[p][i] = -V[p][i]
Step 14: R[p][i] = -R[p][i]
Step 15: End If,For,For
Step 16: For → p to Np //Find fitness for each particles
Step 17: M[p] = fitness(R[p])
Step 18: M[p] = -M[p]
Step 19: End For
Step 20: For → j to Nt
Step 21: For → p to Np
Step 22: For → i to Nd
Step 23: R[p][i] = R[p][i] + V[p][i]
Step 24: IF R[p][i] >xMax
Step 25: R[p][i] = xMax;
Step 26: Else if R[p][i] <xMin
Step 27: R[p][i] = xMin;
Step 28: End if, For, For
Step 29: For → p to Np
Step 30: M[p] = fitness(R[p])
Step 31: M[p] = -M[p];
Step 32: If M[p] >pBestValue[p] // swapping best values and position
Step 33: pBestValue[p] = M[p]
Step 34: For → I to Nd
Step 35: pBestPosition[p][i] = R[p][i]
Step 36: End For, If
Step 37: If M[p] >gBestValue
Step 38: gBestValue = M[p]
Step 39: For I to Nd
Step 40: gBestPosition[i] = R[p][i]
Step 41: End For, If, For
Step 42: bestFitnessHistory[j] = gBestValue
Step 43: w = wMax - ((wMax-wMin)/Nt) * j
Step 44: For → p to Np
Step 45: For → I to Nd
Step 46: r1 & r2 = rand
Step 47: V[p][i] = chi * w * (V[p][i] + r1 * c1 * (pBestPosition[p][i] - R[p][i]) + r2*c2 * (gBestPosition[i] - R[p][i])); // //Fixing best value and position
Step 48: If V[p][i] >vMax
Step 49: V[p][i] = vMax;
Step 50: End If, For, For, For
Step 51: double abs = Math.abs(gBestValue);
Step 52: Achieved Best Fit Value of the matrix.
Step 53: Square root (abs) = sqabs
Step 54: If entries <sqabs// condition to fix entries
Step 55: entries = 0
Step 56: Else
Step 57: entries = 1
Step 58: End If
Step 59: Achieved Binary Matrix = [BMat]
Step 60: BMat as input to again PSO with INV components
Step 61: INV → change R to C & C to R as MBPSO
Step 62: N number of values are achieved. Where N is number of columns
Step 63: N number are arranged in ascending order to get least n number of attributes
Step 64: n number is depends upon the criteria.
Step 65: Achieved effective attributes are selected by MBPSO.

IV. RESULTS AND DISCUSSION

This section demonstrates the performance analysis of the proposed methodology. The performance of this work is evaluated and compared with the existing algorithms.

A. Performance Measures

The performance of the proposed methodology is evaluated using some performance metrics such as execution time, processing time, accuracy and time index. This can be discussed as follows.

Execution time

Execution time is defined as the amount of time required to execute the given process. This can be given as,

$$ExecutionTime = Endingtimeoftheprocess - Startingtimeoftheprocess$$

Processing time

Retrieval Number C4932098319/2019©BEIESP
DOI: 10.35940/ijrte.C4932.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Processing time is represented as the amount of time required to process the given data. This can be given as,

Processing$Time=Endingtimeoftheprocess-Startingtimeoftheprocess$

Accuracy Index

The accuracy index can be defined as the ratio of sum of all the accuracy for entire item set to the maximum accuracy achieved. This can be represented as,

$$Accuracy\ Index = \frac{\sum accuracy}{max(accuracy)}$$

Time index

The time index can be defined as the ratio of total time required for processing entire item set to the maximum time taken.

$$Time\ Index = \frac{\sum time}{max(time)}$$

B. Performance Analysis

The analysis of the performance of this work is described as follows. Here the time taken for execution, processing and accuracy are validated and compared with the tradition algorithms.

Table 1: Execution time for census dataset

Census Dataset	FP-Growth	Éclat	Proposed
30	1.21	1.79	0.75
40	1.16	0.75	0.7
50	0.86	0.72	0.69
60	0.73	0.69	0.6
70	0.68	0.64	0.5

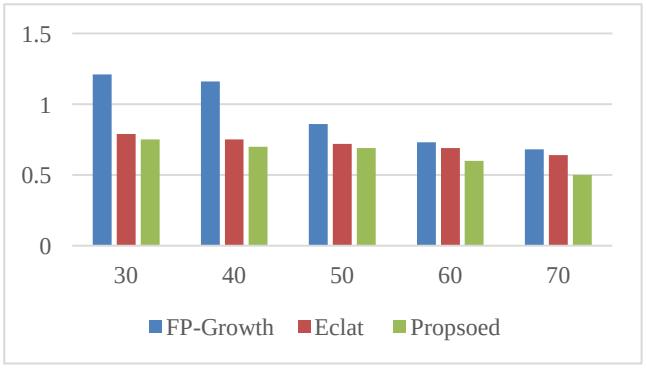

Fig. 2.Execution time for census dataset

Table 1 shows the comparative analysis of the execution time for census dataset. Here various existing algorithms such as FP-Growth and Éclatare used for comparing the performance of the proposed algorithm. From the results it is analyzed that the proposed algorithm requires less time for execution compared to other algorithms. This comparative analysis has been represented graphically in fig 2.

Table 2: Execution time for adult dataset

Adult Dataset	FP-Growth	Éclat	Proposed
30	0.56	0.54	0.5
40	0.5	0.49	0.4
50	0.49	0.45	0.38
60	0.48	0.44	0.3
70	0.42	0.4	0.28

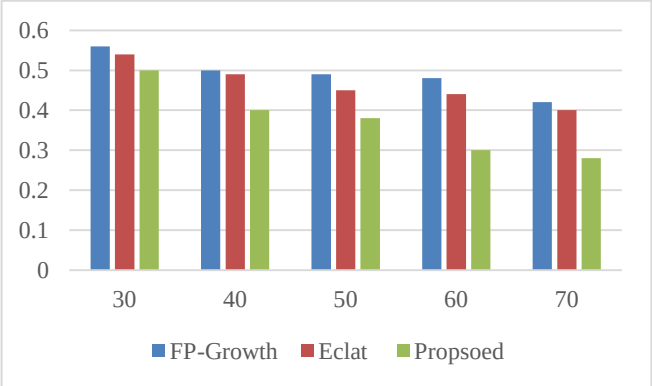

Fig. 3.Execution time for adult dataset

Table 2 depicts the comparative analysis of the execution time for adult dataset. Here various existing algorithms such as FP-Growth and Éclat are used for comparing the performance of the proposed algorithm. From the results it is analyzed that the proposed algorithm requires less time for execution compared to other algorithms. The above records have been represented graphically in Fig 3.

Table 3: Execution time for Letter recognition dataset

Letter Recognition Dataset	FP-Growth	Éclat	Proposed
30	0.21	1.21	0.2
40	0.2	0.21	0.19
50	0.18	0.2	0.15
60	0.17	0.19	0.14
70	0.15	0.17	0.1

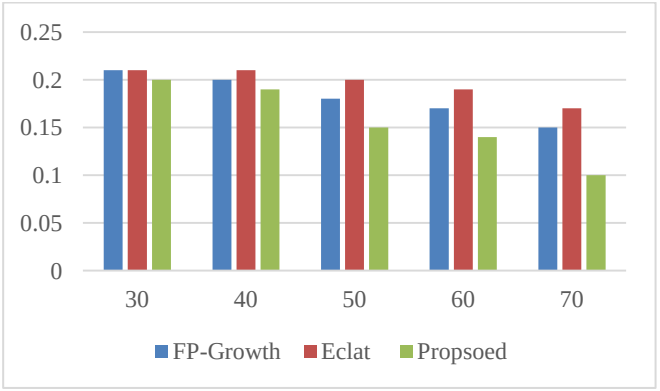

Fig. 4.Execution time for Letter recognition dataset

Table 3 represents the comparative analysis of the execution time for letter recognition dataset. Here various existing algorithms such as FP-Growth and Éclat are used for comparing the performance of the proposed algorithm. From the results it is analyzed that the proposed algorithm requires less time for execution compared to other algorithms. It has been represented graphically in fig 4.

Table 4: Execution time for Mushroom dataset

Mushroom Dataset	FP-Growth	Éclat	Proposed
30	0.13	0.11	0.1
40	0.11	0.11	0.1
50	0.09	0.09	0.08
60	0.08	0.09	0.07
70	0.08	0.08	0.05

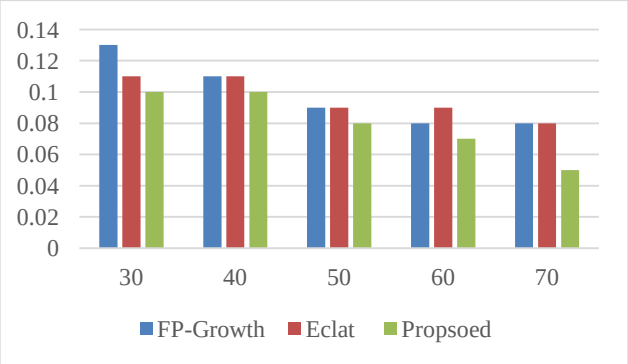

Fig. 5.Execution time for Mushroom dataset

Table 4 shows the comparative analysis of the execution time for mushroom dataset. Here various existing algorithms such as FP-Growth and Éclat are used for comparing the performance of the proposed algorithm. From the results it is analyzed that the proposed algorithm requires less time for execution compared to other algorithms. Fig 5 represents the comparative analysis of execution time for Mushroom dataset using graph.

Table 5: Processing time for various algorithms

Apriori Time (sec)	SAR Time (sec)	Max. Constraints Time(sec)	SARM SMC Time (Sec)	Proposed
97	94	93	91	90
109	105	104	101	100
119	113	112	109	105
134	126	127	122	120
156	142	148	138	130
196	169	181	164	160

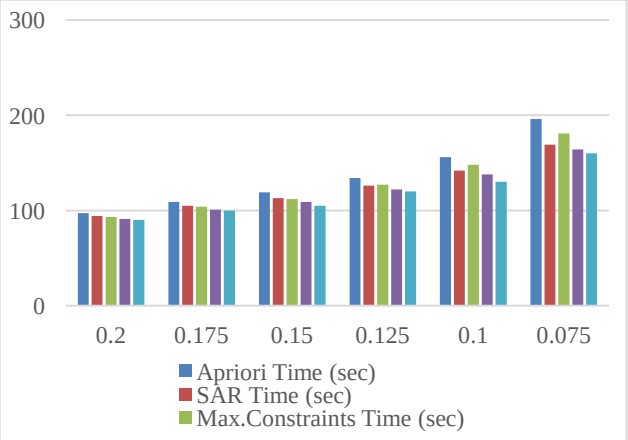

Fig. 6.Processing time for various algorithms

Table 5 shows the comparative analysis of the processing time for different algorithms. Here various existing algorithms such as Apriori, SAR, Maximum constraints and SARM SMC are used for comparing the performance of the proposed algorithm. From the results it is analyzed that the proposed algorithm requires minimum processing time compared to the other algorithm. Fig 6 depicts the graphical representation of processing time on various algorithms.

Table 6: Item set and Rule generation time

Algorithms	Item Generation Time (sec)	Rule Generation Time (sec)
Apriori	137	12
SAR	137	3
Maximum Constraint Time	137	12
SARM SMC	129	3
Proposed	100	2

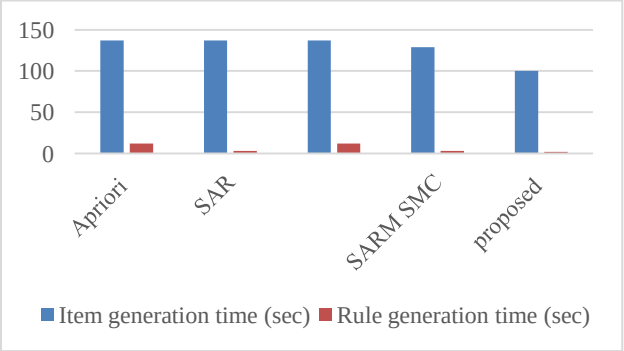

Fig. 7.Item set and Rule generation time

Table 6 displays the comparative analysis of the item set and rule generation time for different algorithms. Here various existing algorithms such as Apriori, SAR, Maximum constraints and SARM SMC are used for comparing the performance of the proposed algorithm. From the results it is analyzed that the proposed algorithm requires minimum time compared to the other algorithm.

Table 7: Accuracy and time index

Algorithms	Time Index%	Accuracy – index %
Apriori	100	97.89
SAR	97.77	100
Maximum Constraint Time	83.445	98.25
SARM SMC	81.4	99.99
Proposed	80	100

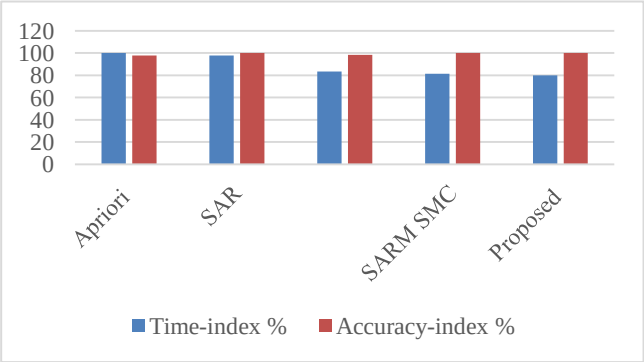

Fig. 8.Accuracy and time index

Table 7 depicts the comparative analysis of the accuracy and time index for different algorithms. Here various existing algorithms such as Apriori, SAR, Maximum constraints and SARM SMC are used for comparing the performance of the proposed algorithm. From the results it is analyzed that the proposed algorithm offers increased accuracy with less time index compared to the other algorithm. The same has been represented graphically in fig 8.

V. CONCLUSION

In this research work, we have represented a novel optimization based ARM algorithm for deriving sensitive and efficient information from the large databases. Initially, a preprocessing procedure has been taken where the input raw data obtained from large database, got converted in binary data type. Secondly, a novel MBPSO algorithm determined the threshold values for minimal support and confidence effectively. Here, fitness function estimation has been done and the best particles are determined. Finally, the most legitimate ARM rules are obtained. In this research work, various experimental and comparative analysis have been done in order to measure the effectiveness of the proposed system. The proposed system has been compared with certain existing methods like FP-Growth, Éclat with respect to processing time for various datasets like Mushroom dataset, letter recognition, etc. From the observations, it has been obvious that the proposed system outperforms other compared methods. Moreover, the proposed MBPSO algorithm has been compared with certain algorithms like Apriori, SAR, SARM SRC, etc. with respect to accuracy-index, time-index, etc. From those analysis, it has been proved that the proposed algorithm has shown remarkable behavior than other algorithms.

REFERENCES

1. F. Feng, J. Cho, W. Pedrycz, H. Fujita, and T. Herawan, "Soft set based association rule mining," *Knowledge-Based Systems*, vol. 111, pp. 268-282, 2016.
2. M. D. Ruiz, J. Gómez-Romero, M. Molina-Solana, J. R. Campaña, and M. J. Martín-Bautista, "Meta-association rules for mining interesting associations in multiple datasets," *Applied Soft Computing*, vol. 49, pp. 212-223, 2016.
3. L. Li, R. Lu, K.-K. R. Choo, A. Datta, and J. Shao, "Privacy-preserving-outsourced association rule mining on vertically partitioned databases," *IEEE Transactions on Information Forensics and Security*, vol. 11, pp. 1847-1861, 2016.
4. M. Al-Maolegi and B. Arkok, "An improved apriori algorithm for association rules," *arXiv preprint arXiv:1403.3948*, 2014.
5. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, pp. 1153-1176, 2016.
6. M. Abdel-Basset, M. Mohamed, F. Smarandache, and V. Chang, "Neutrosophic Association Rule Mining Algorithm for Big Data Analysis," *Symmetry*, vol. 10, p. 106, 2018.
7. D. Martin, A. Rosete, J. Alcala-Fdez, and F. Herrera, "A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules," *IEEE Transactions on Evolutionary Computation*, vol. 18, pp. 54-69, 2014.
8. A. Moustafa, B. Abuelnasr, and M. S. Abougabal, "Efficient mining fuzzy association rules from ubiquitous data streams," *Alexandria Engineering Journal*, vol. 54, pp. 163-174, 2015.
9. D. Nguyen, L. T. Nguyen, B. Vo, and W. Pedrycz, "Efficient mining of class association rules with the item set constraint," *Knowledge-Based Systems*, vol. 103, pp. 73-88, 2016.
10. N. M. Hassan, "Analysis and Implementation some of Data Mining Algorithms by Collecting Algorithm based on Simple Association Rules," *International Journal of Computer Applications*, vol. 138, 2016.
11. H. Jafarzadeh, R. R. Torkashvand, C. Asgari, and A. Amiry, "Provide a new approach for mining fuzzy association rules using apriori algorithm," *Indian Journal of Science and Technology*, vol. 8, pp. 127-134, 2015.
12. V. Beiranvand, M. Mobasher-Kashani, and A. A. Bakar, "Multi-objective PSO algorithm for mining numerical association rules without a priori discretization," *Expert Systems with Applications*, vol. 41, pp. 4259-4273, 2014.
13. G. Qian, C. R. Rao, X. Sun, and Y. Wu, "Boosting association rule mining in large datasets via Gibbs sampling," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 4958-4963, 2016.
14. D. S. da Cunha, R. S. Xavier, D. G. Ferrari, and L. N. de Castro, "Association rule mining using a bacterial colony algorithm," in *Computational Intelligence (LA-CCI), 2015 Latin America Congress on*, 2015, pp. 1-6.
15. B. K. Jassim, A. M. Abdulla, and G. H. Majeed, "Using Genetic Algorithm for Extracting Association Rules," *JOURNAL OF ENGINEERING AND SUSTAINABLE DEVELOPMENT*, vol. 15, pp. 23-30, 2018.
16. M. Bansal, D. Grover, and D. Sharma, "Sensitivity Association Rule Mining using Weight based Fuzzy Logic," *Global Journal of Enterprise Information System*, vol. 9, pp. 1-9, 2017.
17. C.-H. Chen, G.-C. Lan, T.-P. Hong, and S.-B. Lin, "Mining fuzzy temporal association rules by item lifespans," *Applied Soft Computing*, vol. 41, pp. 265-274, 2016.
18. M. Jin, H. Wang, and Q. Zhang, "Association rules redundancy processing algorithm based on hypergraph in data mining," *Cluster Computing*, pp. 1-10, 2018.

AUTHORS PROFILE

Dr.K.Kala is currently working as a Head & Associate Professor in Nachiappa Swamigal Arts and Science College. She has 16 years of academic experience and guided several research scholars in education administration and academia.