Volume 4, Issue 1, January -2017, pp. 4-8 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

Implementation of Optimization Using Eclat and PSO for Efficient Association Rule Mining

M.Sathya[1], Dr.K.Thangadurai[2],

Ph.D Research Scholar[1],

PG and Research Department of Computer Science, Government Arts College (Autonomous), Karur, India,

Email ID: sathyabala333@gmail.com.

Assistant Professor& Head[2],

PG and Research Department of Computer Science, Government Arts College (Autonomous), Karur, India,

Email ID: ktramprasad04@gmail.com.

Abstract:-In this paper, the IEPSO-ARM technique used Eclat algorithm for generating the association rules. With help of Eclat algorithm, IEPSO-ARM technique initially estimates the support value to find the frequent items in the dataset and then determines correlation value to generate the association rules. Finally, the IEPSO-ARM technique designed an Eclat based Particle Swarm Optimization (E-PSO) algorithm for generating the optimized association rule to analyze the frequently buying products by customer in supermarkets and to improve sales growth maintenance of supermarkets. The performance of IEPSO-ARM technique is tested with the metrics such as running time for frequent itemset generation, memory for association rule generation and number of rules generated.

Keywords: frequent item set, Eclat, PSO, association rule mining, supermarkets

1. Introduction

Discovering frequent itemsets is one of the most significant fields of data mining and also identifying the rules associated with the itemset is another essential field in association rule mining. Recent many research works has been developed for efficient association rule mining. For example, an efficient algorithm called FIN was designed in [1] to extract frequent itemsets in databases based on Nodesets. A novel algorithm for mining frequent itemsets in databases was developed in [2] that utilize an efficient pruning strategy called Children–Parent Equivalence pruning to significantly reduce the search space. But, both the above said methods are suffered from the optimized rule generation.

An improved FP tree based Frequent Item set mining algorithm was presented in [3] for association rule mining. The FP tree based Frequent Item set mining algorithm reduces the complex intermediate process of the frequent item set generation and lessens the number of times the transactional database is scanned and remove the generation of the conditional FP trees. Hybrid architecture was designed in [4] by using integrated distributed and parallel computing concept for efficiently discovering frequent item sets from large databases with minimum time. Efficient Association Rule Mining Using Improved Apriori Algorithm was developed in [5] for discovering association rules from large data warehouses. However, the negative association rules remained unsolved.

2. Related Works

An efficient and optimized algorithm called as Customer Purchase Behavior (CPB) was designed in [6] for discovering frequent itemsets by using minimum scans, time and memory and the rules are generated. The CPB algorithm comprises of two parts namely generating frequent itemsets and predicting the customer behavior on purchase of frequent itemsets. However, CPB algorithm takes more time for generating associated rules. A novel method was developed in [7] to generate the association rule mining to identify factors which contribute to heart disease in males and females. But, the risk intensifies with complex and severe heart disorder. An algorithm for mining Frequent Weighted Itemsets (FWI) from weighted items transaction database by using WIT-trees was implemented in [8].

Association Rule Generation for Student Performance Analysis using Apriori Algorithm was designed in [9] to find out the average and below average students and to develop their performance to present better results. The comparative analysis of association rule generation algorithms in data streams was presented in [10]. Though, the number of scans and execution time remained unsolved. In [11], association's rules generation by using Apriori and FPGrowth algorithms was explained to discover frequent itemsets without using candidate key generations, hence improving performance. A novel positive and negative association rule mining algorithm was presented in [12] with aid of improved frequent pattern tree for association rules.

3. Proposed Methodology

With the help of above stated association rule, we can suppose that in the future, those who purchase baby soap are most likely to purchase baby lotion. Such information assists the retailers to discover chances for cross-selling and thus the sale growth from the angle of sales in super market is improved. Let us consider '$item_i$' contains the set of items. A dataset 'DS' over '$item_i = item_1, item_2, \dots, item_n$' signifies the item set and '$trans_i = trans_1, trans_2, \dots, trans_n$' represents the set of transactions which contains items from '$item_i$'. Then a transaction '$Trans$' over '$item_i$' is a couple '$(trans_{id}\, , item_i)$' where '$trans_{id}$' indicates the unique transaction identifier whereas '$item_i$' denotes the set of items from the itemset. The association rule is formulated as follows.

$$t(A) = \{trans_{id} | (trans_{id}, item_i) \in trans_i, a \subseteq item_i \quad (1)$$

Eclat Algorithm for Association Rule Generation

The IEPSO-ARM) technique used Eclat algorithm for finding the frequent items in the given dataset. With the support of the vertical layout format, Eclat algorithm instead of explicitly listing all transactions employs an intersection property to estimate the support of an itemset. In this manner, the support of an itemset 'A' is easily determined by simply intersecting '$trans_{id}$' of any two subsets '$M, N \in A$' such that '$M \cup N = A$'. Subsequently, support of an itemset 'A' in '$trans_i$' represents the cardinality of it '$trans_{id}$'. This entails that support of 'A' is the number of transactions containing 'A' in '$trans_i$' which is mathematically formulated as,

$$\sup(A) = trans_i\,(A) \quad (2)$$

Let us assume a minimum support threshold value to be '$support_{min}$'. Then an itemset 'A' is said to be frequent if its support is not less than a predefined minimum support threshold '$support_{min}$'. Given a transaction dataset 'DS' over a training dataset 'TD' and a minimal support threshold '$support_{min}$', then set of frequent itemsets is formulated as,

$$FI\,(trans_i, support_{min}) = \{A \subseteq D \mid \sup(A) \geq support_{min}\}$$
(3)

During the first scan of dataset 'DS', for each single item a '$trans_{id}$'' is maintained and items are extracted with the list of number of transactions in which they are present with the help of intersection property which formulated as follows.

$$t(PM) \cap d\,(PN) = \big(\big(t(P) \cap t(M)\big) \cap (t(P) - t\,(N)) \\ (t(P) \cap t) - \big(t(P) \cap t(M) \cap t(N)\big) t(PM) - t(PMN) = d\,(PMN)$$
(4)

Thus, '$n+1$' itemset are produced from 'n' itemset by means of performing intersection of '$trans_{id}$' of frequent 'n' item set. This process is continual until no candidate itemsets are found. After generating the frequent itemsets, a correlation measure called '$Lift$' is employed to generate the association rule which is mathematically formulated as follows,

$$Lift\,(X, Y) = \frac{Prob\,(X \cup Y)}{Prob\,(X) * Prob\,(Y)}$$

From the equation (4), interesting association rules about purchasing behavior of customer is determined where X, Y represents the two itemsets. The Eclat Algorithm for Association Rule Generation is shown below Figure 1.

Eclat based Particle Swarm Optimization (E-PSO) for Optimized Rule Generation

Each particle in E-PSO Algorithm has a 'position' and 'velocity' whereas position is represented as a solution suggested by the particle. Velocity is the rate of changes of the next position with respect to current position. The position and velocity values are arbitrarily initialized in E-PSO Algorithm. E-PSO Algorithm contains collection of random particles i.e. rules. During the each iteration, all particles are updated by using $pbest$ and$gbest$ values. $pbest$ represents the best solution it has achieved so far. Another best value is $gbest$ acquired so far with any particle in the population. Afterward, the particle updates its velocity by using following mathematical formula,

$$vel_i = vel_i + c1 * rand() * (pbest[] - p_i) + c2 * rand() * (gbest[] - pos_i)$$
(6)

At the each iteration, the position of particle is updated by using following mathematical formula

$$pos_i = pos_i + vel_i \quad (7)$$

From the equation (6), (7), i denotes s a particles number i.e. $i = 1, .., N$, whereas N indicates the number of particles in the swarm. vel_i Represents the particle velocity, pos_i is the position of current particle and rand () denotes a random number between (0, 1). Here c1, c2 are learning factors usually which are assigned as c1 = c2 = 2. In E-PSO Algorithm, the fitness value is employed to estimate the importance of each particle i.e. rules. The fitness value is measured by using the fitness function. The fitness value of any rule depends on its support and confidence values which is mathematically formulated as follows

$$Fit(k) = Con(k) \times [log(Sup(k) \times Length(k) + 1) \quad (8)$$

From the equation (8), $Fit(k)$ is the fitness value of rule type k, $Con(k)$ is the confidence of rule type k, $Sup(k)$ is the actual support value of rule type k and $Length(k)$ is the length of rule type k. E-PSO Algorithm want to maximize this fitness function. Larger support and confidence values result in a greater association, which represents the rule is significant. The particle in the population that has the highest fitness value is selected as "gbest," and its support and confidence are employed as the minimum thresholds.

// E-PSO Algorithm for Optimized Rule Generation
Input: output of Eclat Algorithm (i.e. generated association rules)
Output: optimized rule (i.e. interesting rules for the organization)
Step 1:Begin
Step 2: Initialize particle with random position and velocities
Step 3: Do
Step 4: **For** each particle
Step 5: Compute fitness value using (8)
Step 6: If the fitness value is better than the best fitness value ($pBest$)
Step 7: Set current value as the new $pBest$
Step 8: **End for**
Step 9: Select the particle with the best fitness value of all the particles as the $gBest$
Step 10: **For** each particle
Step 11: Update particle velocity using (6)
Step 12: Update particle position using (7)
Step 13: **End for**
Step 14: While maximum iterations or minimum error criteria is not attained
Step 15:End while
Step 16:End

Figure 1 E-PSO Algorithm for Optimized Rule Generation

As shown in Figure 1, E-PSO algorithm initially takes the association rules generated from the Eclat Algorithm as input. Then, it initializes the particles with random position and velocity with the aid of generated association rules where each particle represents rules. With the support of determined fitness value, the particle with the highest fitness value is selected as the $gBest$.

4. Result and discussions

In this section, the result analysis of IEPSO-ARM technique is evaluated. The performance of IEPSO-ARM technique is compared against with exiting two methods namely, Frequent Item Sets using Nodesets (FIS-N) [1], Frequent Item Sets using Equivalence Pruning (FIS-EP) [2] and Eclat based ant colony optimization (E-ACO) algorithm. The performance of IEPSO-ARM technique is evaluated along with the following metrics with the aid of tables and graphs.

Measurement of Running Time for Frequent Itemset Generation

The running time for frequent itemset generation is measured in terms of milliseconds (ms) and mathematically formulated as follows,

$$\text{running time} = n * \text{time(generating one frequent itemset)}$$

From the equation (9), running time for frequent itemset generation is obtained where n represents the number of frequent itemsets generated. When the running time for frequent itemset generation is lower, the method is said to be more efficient.

Support Value	Running Time for Frequent Itemset Generation (ms)			
	FIS-N	FIS-EP	E-ACO	IEPSO-ARM technique
0.1	190	171	142	138
0.2	168	156	125	115
0.3	145	138	109	97
0.4	120	113	95	82
0.5	101	95	81	75

Table 1 show the result analysis of running time

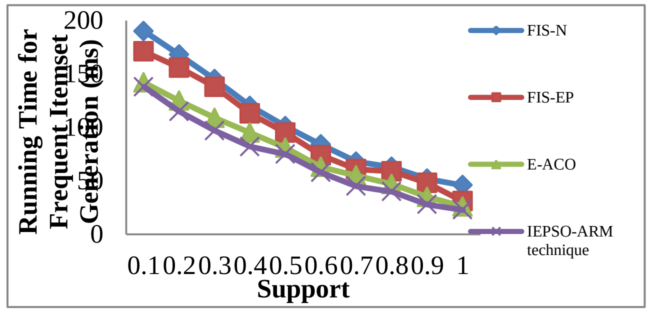

Figure 2 Measure of Running Time for Frequent Itemset Generation

Therefore, proposed IEPSO-ARM technique reduces the running time of frequent itemset generation by 55% as compared to FIS-N [1] and 38% as compared to FIS-EP [2] and 13% as compared to E-ACO algorithm respectively.

Measurement of Memory Consumption for Association Rule Generation

The memory consumption for association rule generation is measured in terms of Megabytes (MB) and mathematically formulated as below,

$$\text{memory consumption} = \text{n} * \text{memory(generating one association rule)} \quad (10)$$

From the equation (10), memory consumption for association rule generation is obtained where n represents the number of association rules generated.

Support Value	Memory Consumption for Association Rule Generation (MB)			
	FIS-N	FIS-EP	E-ACO	IEPSO-ARM technique
0.1	25.5	23.8	20.1	18.3
0.2	23.1	21.5	18.5	16.9
0.3	20.3	19.3	16.9	15.3
0.4	17.5	16.7	15.6	14.2
0.5	16.4	15.2	14.1	13

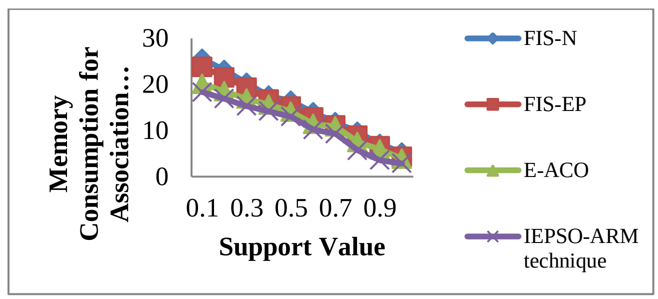

Figure 3: Measure of Memory for Association Rule Generation

This in turn assists to reduce the memory consumption of association rule generation in a significant manner. Therefore, Proposed IEPSO-ARM technique reduces the memory consumption of association rule generation by 47% as compared to FIS-N [1] and 35% as compared to FIS-EP [2] and 21% as compared to E-ACO algorithm respectively.

Measurement of Number of Rules Generated

In IEPSO-ARM technique, the Eclat algorithm is used for generating the number of rules based on support value and correlation measure. When the number of rules generated is lower, the method is said to be more efficient.

Support Value	Number of Rules Generated			
	FIS-N	FIS-EP	E-ACO	IEPSO-ARM technique
0.1	2250	1700	1380	1250
0.2	2080	1510	1240	1140
0.3	1800	1350	1130	1020
0.4	1790	1120	1030	950
0.5	1610	950	870	864

Table 3 Tabulation for Number of Rules Generated

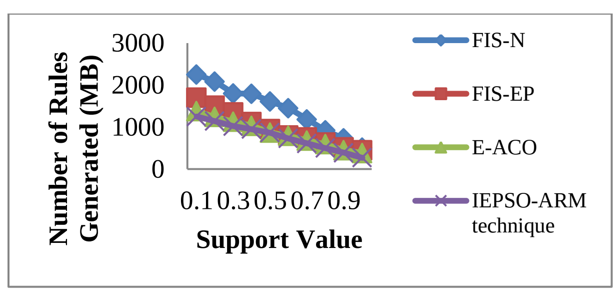

Figure 4 Measure of Number of Rules Generated

This in turn supports to reduce the number of rules generated in an efficient manner. As a result, proposed IEPSO-ARM technique reduces the number of rules generated by 87% as compared to FIS-N [1] and 29% as compared to FIS-EP [2] and 13% as compared to E-ACO algorithm respectively.

5. Conclusion

In this work, An Integrated Eclat and PSO based Association Rule Mining (IEPSO-ARM) technique is developed for minimizing the frequent itemset and association rule generation and therefore improve the point of sales in supermarkets. In contrast to existing methods, IEPSO-ARM technique utilize vertical layout format to optimize the rules generated based on the correlation measure, i.e. rules are not only generated based on the correlation measure but also on the intensity of correlation measure. The efficacy of IEPSO-ARM technique is measured in terms of running time for frequent itemset generation, memory for association rule generation and number of rules generated using Sample Dataset for Market Basket Analysis and compared with existing methods.

Acknowledgement

We would like to thank our Principal, HOD and faculty members of P.G & Research Department of Computer Science & Research Scholar, Government Arts College, Karur.

6. References

[1] Zhi-Hong Deng, Sheng-Long Lv, "Fast mining frequent itemsets using Nodesets", Expert Systems with Applications, Elsevier, Expert Systems with Applications, Elsevier, Volume 41, Pages 4505–4512, 2014

[2] Zhi-Hong Deng, Sheng-Long Lv, "PrePost+: An efficient N-lists-based algorithm for mining frequent itemsets via Children–Parent Equivalence pruning", Expert Systems with Applications, Elsevier, Volume 42, Pages 5424–5432, 2015

[3] Meera Narvekara, Shafaque Fatma Syed, "An optimized algorithm for association rule mining using FP tree", Procedia Computer Science, Elsevier, Volume 45, Pages 101 – 110, 2015

[4] Anil Vasoya, Dr. Nitin Koli, "Mining of association rules on large database using distributed and parallel computing", Procedia Computer Science, Elsevier, Volume 79, Pages 221 – 230, 2016

[5] Ish Nath Jha, Samarjeet Borah, "Efficient Association Rule Mining Using Improved Apriori Algorithm", International Journal of Scientific & Engineering Research, Volume 3, Issue 11, Pages 1-4, November-2012

[6] Manali Rajeev Raut, Hemlata Dakhore, "An Approach to Mining Association Rules in Horizontally Distributed Databases with Anonymous ID Assignment", IEEE 2015 Global Conference on Communication Technologies (GCCT), Pages 23-24, April 2015

[7] M. Krishnamurthy, E. Rajalakshmi, R. Baskaran, A. Kannan, "Prediction of customer buying nature from frequent itemsets generation using Quine-McCluskey method", IET Chennai Fourth International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2013), Pages 12-14, 2013

[8] Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle, Yi-Ping Phoebe Chen, "Association rule mining to detect factors which contribute to heart disease in males and females", Expert Systems with Applications, Elsevier, Volume 40, Pages 1086–1093, 2013

[9] Bay VO, Frans Coenen, Bac Le, "new method for mining Frequent Weighted Itemsets based on WIT-trees", Expert Systems with Applications, Elsevier, Expert Systems with Applications, Elsevier, Volume 40, Pages 1256–1264, 2013

[10] D. Magdalene Delighta Angeline, "Association Rule Generation for Student Performance Analysis using Apriori Algorithm", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Volume 1, Issue 1, Pages 12-16, March-April 2013

[11] Dr. S. Vijayarani and Ms. R. Prasannalakshmi, "Comparative Analysis of Association Rule Generation Algorithms in Data Streams", International Journal on Cybernetics & Informatics (IJCI), Volume 4, Issue 1, Pages 15-25, February 2015

[12] J.Suresh, P.Rushyanth,Ch.Trinath, "Generating associations rule mining using Apriori and FPGrowth Algorithms", International Journal of Computer Trends and Technology (IJCTT), volume4, Issue4, Pages 887-892, April 2013

[13] Ruchi Bhargava, Prof. Shrikant Lade, "Effective Positive Negative Association Rule Mining Using Improved Frequent Pattern", International Journal of Modern Engineering Research (IJMER), Volume 3, Issue 2, Pages 1256-1262, March-April. 2013

[14] Sample Dataset for Market Basket Analysis:http://recsyswiki.com/wiki/Grocery_shopping_datasets