

**OPTIMAL PATTERN MINING USING DIFFERENTIAL EVOLUTION
ALGORITHM**¹Priyanka Gupta, ²Akhilesh Tiwari¹Department of CSE & IT, Madhav Institute of Technology & Science, Gwalior, M.P, India²Department of CSE & IT, Madhav Institute of Technology & Science, Gwalior, M.P, India

ABSTRACT :- At present time, data are rapidly grown day by day. There is an urgent need of storage of data in data warehouse in an efficient manner. But sometime, there are some patterns which are not optimal but store as frequent patterns. These patterns are not useful most of the time so there is need to get optimal pattern from frequent patterns. Now, this is the big challenge for researcher to find most optimal patterns. For working on this issue apply differential evolution optimization technique to get optimal pattern out of generated frequent patterns from apriori algorithm. Experimental results shows that apriori with DE algorithm gives most optimal patterns in reduced time rather than Genetic Approach.

KEYWORDS:- DataMining, Optimal pattern mining, Differential Evolution algorithm, Apriori algorithm.

INTRODUCTION

Data Mining alludes to the way toward separating efficient, updated, latent, helpful & the understandable pattern from a huge incomplete, noise, non-stable and random data [1]. An important subfield of data mining that is called pattern mining. Pattern mining comprises of creating DM algo to find exciting, unpredicted & valuable patterns in dataset. Pattern mining algo should be applied on different sorts of information for example, transaction databases, sequence databases, streams, strings, spatial datagraphs. Pattern mining algo can be intended to find different kinds of patterns for example, sub-graphs, associations, indirect associations, trends, periodic patterns, sequential rules, lattices, sequential patterns & high-utility patterns. But what is an interesting pattern? There are a few definitions. For instance, few scientists characterize a fascinating pattern as a pattern that shows frequently in a dataset. Different analyst needs to find uncommon patterns, patterns with a high confidence, the top patterns and so on. [2].

Literature Survey

Charles X. Ling et al [2002] Illustrate a novel algo that proposes activities to change clients from an undesired status (for example, attritors) to a preferred one (for example,loyal). Their algorithm considers the cost of activities & further it endeavors to augment the normal net benefit expected net profit. To our best knowledge, no DM algo or apparatus today can achieve his significant task in CRM. The algo is executed, with many propelled features in a specific & exceptionally effective DM software known as proactive solution [3].

Xilu Wang and Weili Yao [2007] handles the sequential mode mining from the viewpoint of system and an optimization method is constructed to aid find the interesting consistent rules. The numerical experiment verifies the validity and efficiency of the algorithm [4].

Mr. M. Saravanan and Dr. V. L. Jyothi [2014] discover sequential patterns of web files using genetic algorithm. This approach can be utilized to examine the recent visitor's trend & prompt the formation of frequent & most visited pages. The reason for utilizing hereditary algo is to discover optimal sequential web pages [5].

Chaoqun Weng and Junsong Yuan [2015] propose a novel branch & bound search based co-occurrence feature mining algo that can specifically mine both optimal conjunctions (AND) & disjunctions (OR) of individual features at arbitrary orders concurrently. This feature mining procedure is co-ordinated into a multi-class boosting structure Adaboost.MH with the end goal that the weighted training fault is limited by the found co-occurrence features in every boosting stage.

Trials on UCI benchmark databanks, the scene acknowledgment dataset & the action recognition dataset approve both the adequacy & proficiency of our proposed method [6].

Willy Ugarte et al [2015] Present the idea of optimal patterns (OPs), characterized as the best patterns as per a given client inclination & demonstrate that OPs include numerous DM issues. At that point, they propose a non-exclusive strategy in view of a dynamic constraint satisfaction problem to mine OPs & we demonstrate that any OP is portrayed by a fundamental constraints & an arrangement of constraints to be progressively included. At last, we present an investigational learning comparing our approach versus ad hoc techniques on a few kind of Ops [7].

Ashok Kmar D and T. A. Usha [2015] Examine all the frequent item-sets & furthermore produces the association rules from the huge databanks via utilizing the proposed genetic algo. In this paper, the apriori algo that arrangements the association rule mining alongside the genetic algo is investigated. The output reveals that time complexity is more and it gets reduced only by using Floating point data, where as memory space occupied is less in Genetic Algorithm (GA) when contrasted to apriori algo [8].

Vishakha Agarwal and Akhilesh Tiwari [2016] Introduce a new pattern mining algo for discover optimal patterns from pattern repository. The algorithm uses the features and operators of the genetic algorithm to incorporate the property of optimality among patterns [9].

PROPOSED METHODOLOGY

A. Problem Statement

- 1) The technique for genetic Algorithm is more composite to calculate. The thought behind the strategy for differential advancement is that the distinction between two vectors yields a distinction vector which can be utilized with a scaling element to navigate the inquiry space. As in the genetic algo starting an irregular population is picked, similarly over the issue space & to make the subsequent generation an equivalent number of donor vectors are produced.
- 2) The power of a genetic algo lies in its capacity to discover a efficient resolution for an issue where the iterative resolution is excessively restrictive in time & the arithmetic solution is not attainable.
- 3) Genetic algorithm are not more effective and more efficient & when operated in algebraic optimization, it is not to be capable to present things as real numbers instead of having to working your way around to a chromosomal kind of representation, first.
- 4) There was the problem with the existing approach is that it generated maximal optimum pattern due to local optima search. It can't achieve their comprehensive resolution.

B. Apriori Algorithm

The issue of frequent pattern mining is to locate the entire arrangement of patterns fulfill a minimum support in the transaction database. The downward closure property is utilized to prune the uncommon patterns. This property express that if a pattern is infrequent, at that point the majority of its super-patterns should be infrequent. The apriori algo is the primary resolution for the frequent pattern mining issue. For the initial dataset scan, apriori discovers entire one element frequent item-sets & depend on that, produces the contender for two-component frequent item-sets. In the next dataset examine, apriori discovers the majority of two-component frequent item-sets & depend on that, creates the possibility for three-component visit item-sets et cetera [10].

C. Differential Evolution Algorithm

Differential Evolution is a stochastic direct search & global optimization algo & is an illustration of an evolutionary Algo from the field of evolutionary computation.

Let the population be signified by P with N people. Every person is a n - dimensional vector. We will utilize t to signify the invention variable. If there is no uncertainty, we will discard the sign t.

$$P(t) = \{x_1^t, x_2^t, x_3^t, \dots, x_N^t\}$$

$$x_i^t = \{\zeta_{1,i}^t, \zeta_{2,i}^t, \zeta_{3,i}^t, \dots, \zeta_N^t\}, \quad i=1, 2, 3, \dots, N$$

DE produces novel offsprings by framing a noisy replica (preliminary vector) of every parent singular (target vector) of the population. The population is effectively enhanced by three fundamental administrators: crossover, selection and mutation. First the transformation administrator makes mutant vectors by bothering each objective vector with the weighted distinction of two other randomly chosen persons. Then, hybrid administrator creates preliminary vectors by blending the parameters of the mutant vectors with the objective vectors as per a chosen likelihood conveyance. Lastly, the selection administrator forms the next generation via deterministically choosing between the preliminary & relating target vector which fits better the objective function [11].

PROPOSED ALGORITHM:

Input : A transactional dataset D.

Output: A set of optimal pattern.

Step 1. Select the Dataset.

Step 2. Apply Apriori algorithm to generate the frequent patterns.

Step 3. Now, initialize generated frequent patterns as a population of size of N-Dimensional vector.

Step 4. Calculate Support Count.

Step 5. Calculate Fitness Function by using formula

Fitness = Support count * Probability * chromosome

1) If fitness value less than one (support count) then discard the population.

Step 6. Apply mutation to create mutant vector

1) Select 3-vectors randomly from population.

$$V = V_1 + \alpha (V_2 - V_3)$$

Here, α is multiplier [0,1]

V is mutant vector

V_1, V_2 and V_3 are Random vectors

Step 7. Apply intersection with possibility 95%

1) Produce preliminary vector via intersection of mutant vector & target vector.

2) Produce a irregular number from arbitrary number producer to check the intersection possibility.

3) If random number is not as much as intersection possibility after that select preliminary vector otherwise restore preliminary to target vector.

Step 8. This process will continue until criteria not met.

Step 9. Stop.

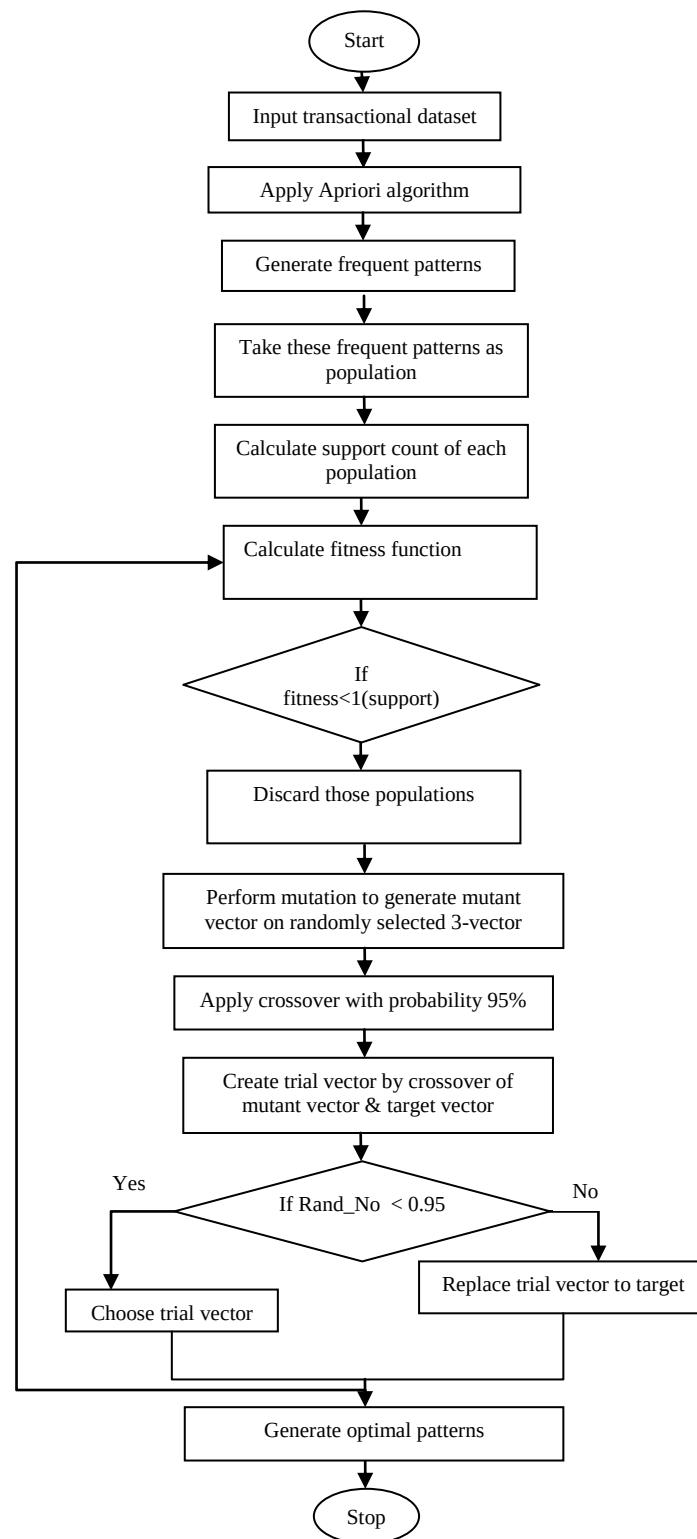


Fig. 3.1 Flowchart of Proposed Work

Proposed algorithm discussed by an example. Consider six transactions as a sample dataset including six items i.e. A, B, C, D, E and F in table 1. Now process it by applying apriori algorithm with 50% support value and extract frequent pattern in table 2.

Table 1. Sample Dataset

TID	Itemset
1	A,B,C,D,E
2	A,B,C,D,F
3	A,B,C
4	A,B,D
5	A,C,D
6	B,C,D,E

Table 2. Frequent Pattern-set

Pattern_Id	Pattern	Supp_Count
P1	A	5
P2	B	5
P3	C	5
P4	D	5
P5	A,B	4
P6	A,C	4
P7	A,D	4
P8	B,C	4
P9	B,D	4
P10	C,D	4
P11	A,B,C	3
P12	A,B,D	3
P13	A,C,D	3
P14	B,C,D	3

Now, the pattern-set shown in Table 2 is given as an input to the algorithm DE along with few other inputs as crossover probability is taken as 95% and mutation probability is taken as 1%. The algorithm works as follows-

Probability- [A-0.5, B-0.25, C-0.1667, D-0.125, E-0.1, F-0.035]

Calculate

Fitness Function = Support* Probability * chromosome

For eg.

$F(A) = 5 * 0.5 * 1 = 0.25$

Consider only those patterns that have value greater than 1.

[P2, P5, P6, P8, P9, P10, P11, P12, P13, P14]

SELECTION –

In the selection process, the three patterns are randomly selected. For eg.

P5- A,B (110000)

P6- A,C (011000)

P8- B,C (101000)

Mutation-

These selected vectors are mutated and generate a mutant vector by using

$$V = V_1 + \alpha (V_2 - V_3)$$

For eg.

$$V = 12 + 1(6 - 10) = 8 \text{ (100000)}$$

Crossover-

Now, crossover operator is applied with 95% probability over target and mutant vector patterns and trial vector is generated. For eg.

P6- A, C (Target vector)

P1- A (Mutant vector)

Now check the condition and if random value is less than crossover probability then choose trial vector otherwise replace trial to target vector and further tests for their fitness in the new population. If the patterns pass their fitness value these are placed in the new population of optimal patterns. Moreover, same process is repeated for other set of vectors patterns. Table 3 shows the optimal patterns extracted out of the old population of pattern-set.

Table 3. Optimal Pattern-SET

Pattern_Id	Itemset
P1	A
P2	B
P3	C
P6	A,C
P14	B,C,D

IV. RESULT ANALYSIS

In the outcome examination, the analysis of proposed work performed by utilizing MATLAB apparatus. Transactional dataset used for the implementation. This firstly loaded dataset then choose minimum support count. Finally get all frequent patterns through apriori algorithm then by applying DE algorithm get most optimal pattern. As a example get 5 optimal patterns out of 14 generated frequent patterns.

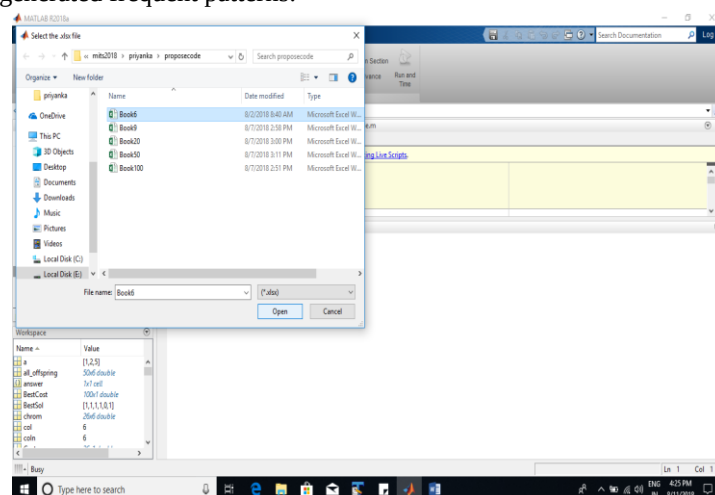


Fig. 4.1 Load dataset

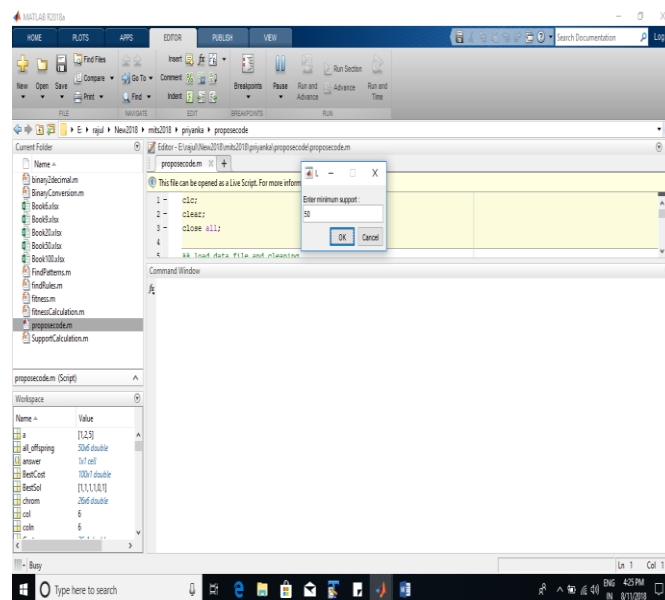


Fig. 4.2 : Input support value

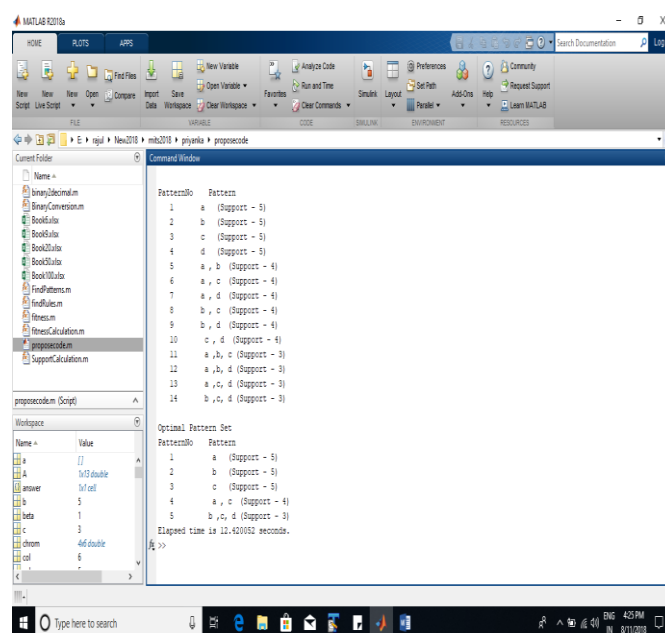


Fig. 4.3 Generated optimal patterns

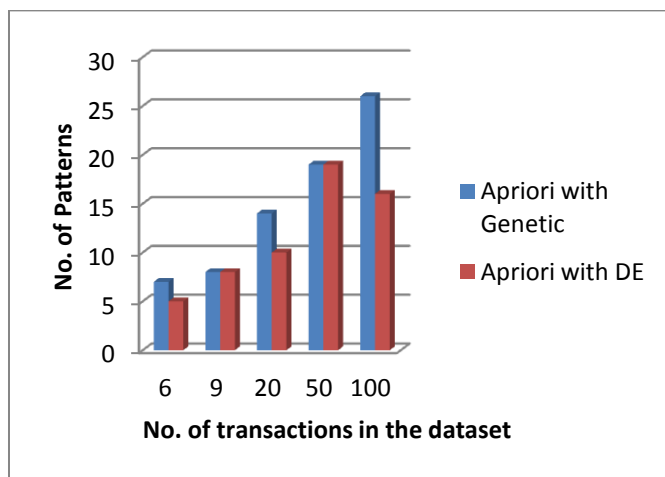


Fig. 4.4 Comparison graph of generated optimal pattern with existing approach over a different no. of transactions

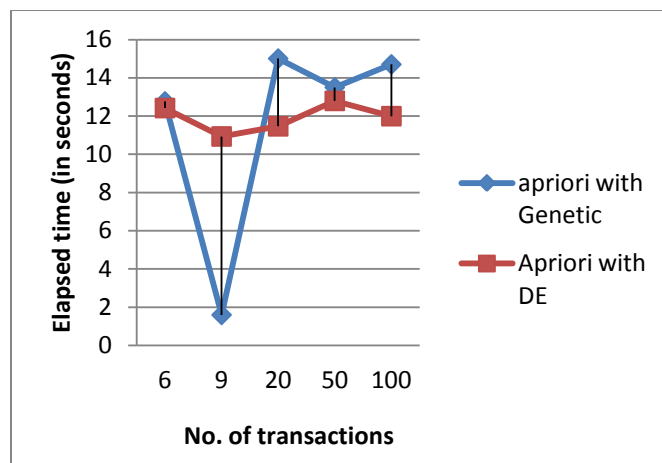


Fig. 4.5 Time taken to generate optimal patterns

CONCLUSION

Pattern warehousing is a new concept, very few researchers has looked into this concept of making the patterns non-unpredictable and store them in persevering way. In the proposed advance firstly locate the frequent patterns & after that store these examples in design stockroom effectively. The algo produces & stores ideal examples which helps in the further decrease of the extent of the archive. The inclusion of DE algorithm provides the strength to the proposed approach. Transactional dataset used for this evaluation. The focus was to get most optimal patterns and also achieve less execution time to produce optimal patterns than the previous algorithm. The model produces the most optimal patterns from the frequent patterns.

References

- [1] Panchal Mr. Krunal and Patel Jaina, 2015. Effective Intrusion Detection System using Data Mining Technique. Journal of Emerging Technologies and Innovative Research (JETIR), Vol. 2, Issue 6, PP. 1869-1878.
- [2] Philippe Fournier-Viger, "An Introduction to Frequent Pattern Mining", Data Mining Blog, 2013.
- [3] Charles X. Ling et al, "Mining Optimal Actions for Profitable CRM", Proceedings IEEE International Conference on Data Mining, 2002, pp. 767-770.
- [4] Xilu Wang and Weili Yao, "Sequential Pattern Mining: Optimum Maximum Sequential Patterns and Consistent Sequential Patterns", Proceedings of the 2007 IEEE International Conference on Integration Technology March 20 - 24, 2007, Shenzhen, China, pp. 365-368.
- [5] Mr. M. Saravanan and Dr. V. L. Jyothi, "A Novel Approach for Sequential Pattern Mining By Using Genetic Algorithm", International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 284-288.
- [6] Chaoqun Weng and Junsong Yuan, "Efficient Mining of Optimal AND/OR Patterns for Visual Recognition", IEEE Transactions On Multimedia, Vol. 17, No. 5, May 2015, pp. 626-635.
- [7] Willy Ugarte et al, "Modeling and Mining Optimal Patterns using Dynamic CSP", IEEE 27th International Conference on Tools with Artificial Intelligence, 2015, pp. 33-40.
- [8] Ashok Kmar D And T. A. Usha, "Improved Apriori Algorithm Using Genetic Algorithm For Itemset Mining", International Journal Of Applied Engineering, Vol. 10, No. 82, December, 2015, pp. 8-12.
- [9] Vishakha Agarwal and Akhilesh Tiwari, "A Novel Optimal Pattern Mining Algorithm using Genetic Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 144 – No.4, June 2016, pp. 18-21.
- [10] Ahmed et al, "Efficient Tree Structures For High Utility Pattern Mining in Incremental Databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 12, December 2009, pp. 1708-1721.
- [11] Xiyu Liu et al, "Mobile Clustering Agents based on Differential Evolution", Third International Conference on Pervasive Computing and Applications, 2008, pp. 9-12.